The Evolution of Human AIDS Viruses

WA YANG

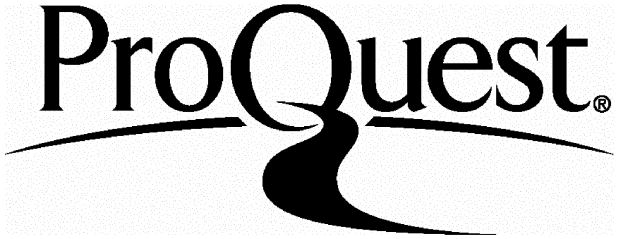
Department of Biology

University College London

2003

Submitted to the University of London

For the Degree of Doctor of Philosophy



ABSTRACT

I investigated variable selective pressures among amino acid sites in HIV-1 genes. Selective pressure at the amino acid level was measured using the nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$). Likelihood ratio tests detected positive selection in every gene in the genome, with the majority located in gp160. Most HIV-1 genes were evolving at a subtype specific manner. As adaptive evolution is driven chiefly by immune detection, this change of selective constraint was indicative of variations in immune targeting. Differences in selective pressure contributed to the extensive genetic diversity observed across the genome and could be viewed as co-evolution of the subtypes.

I measured the physiochemical properties of amino acids and found that those at positive selection sites were more diverse than those at variable sites. Furthermore, amino acid residues at exposed positive selection sites were more physiochemically diverse than at buried positive selection sites.

I also examined the evolution of HIV-2 and SIVmac after the cross-species transmissions. My results indicated that HIV-2 did not appear to be evolving at a faster rate than the progenitor lineages. It is possible that fewer adaptive changes in the progenitor virus were required for successful infection. A notably different pattern was observed for SIVmac lineages. My findings showed that SIVmac lineages appeared to be evolving much faster than HIV-2. Also a fraction of sites in SIVmac lineages that were evolving by relaxed functional constraint became positively selected post zoonosis.

CONTENTS

ABSTRACT 2

CONTENTS 3

LIST OF TABLES 7

LIST OF FIGURES 9

ACKNOWLEDGEMENTS 10

INTRODUCTION 11

CHAPTER 1 THE EVOLUTION AND ORIGIN OF PRIMATE LENTIVIRUS 15

1.1 THE GENOMIC ORGANISATION AND INFECTION MECHANISM OF PRIMATE LENTIVIRUSES 16

1.1.1 The Genetics and the Morphology of Primate Lentivirus 16

1.1.2 The Life Cycle of Primate Lentiviruses 18

1.1.3 The Virulence 21

1.2 THE EVOLUTION OF NON-HUMAN PRIMATE LENTIVIRUSES 23

1.2.1 The Origin and the Diversity of Simian Immunodeficiency Viruses 23

1.2.2 Selection and Recombination 26

1.3 THE ORIGIN OF HUMAN AIDS VIRUSES 29

1.3.1 The Origin of HIV-2 29

1.3.2 The Origin of HIV-1 32

1.4 THE DIVERSITY AND EVOLUTION OF HUMAN AIDS VIRUSES 34

1.4.1 The Divergence of HIV-2 groups 34

1.4.2 The Genetic Diversity of HIV-1 35

CHAPTER 2 EXTENSIVE ADAPTIVE EVOLUTION DETECTED IN THE HUMAN IMMUNODEFICIENCY VIRUS TYPE I GENOME 37

2.1 THE HOST IMMUNE SYSTEM AND HIV-1 38

2.2 THE CROSS-SECTIONAL STATISTICAL STUDY 40

2.2.1 Dataset and Phylogenetic Inference 40

2.2.2 Estimation of d_N/d_S (ω) Ratios Across the Genome 40

2.2.3 Likelihood Ratio Tests and Bayesian Inference 41

2.2.4 Amino Acid Acceptability, Protein 3D Structure and Epitope Mapping 42

2.3 RESULTS 43

2.3.1 Positive Selection and HIV-1 Genes 43

2.3.2 Amino Acid Diversity, Protein Tertiary Structure and Immunogenic Epitopes 47

2.4 DISCUSSION 52

2.4.1 Evidence of Adaptive Evolution in the HIV-1 Genome 52

2.4.2 The Influence of Recombination 54

2.4.3 Amino Acid Substitution Patterns at Positive Selection Sites 55

2.4.4 Diversifying Selection, Antigenic Variation and Epitope Evolution 59

CHAPTER 3 PATTERNS OF EVOLUTION OBSERVED IN HIV-1 62

3.1 SUBTYPE-SPECIFIC VARIATION IN SELECTIVE PRESSURE 63

3.2 AMINO ACID SUBSTITUTIONS: CONSERVATIVE VERSUS RADICAL 63

3.3 DATA AND MODELS 66

3.3.1 Sequence Data and Phylogeny Inferences 66

3.3.2 Detecting Positive Selection in HIV-1 Genome 67

3.3.3 Amino Acid Substitution Models 68

3.4 RESULTS 69

3.4.1 Adaptive Evolution Operating at Different Sites in Different Subtypes 69

3.4.2 Conservative Amino Acid Substitutions Fixed at a Higher Rate 72
3.5 DISCUSSION 75
3.5.1 Evidence of Long Term Recurrent Selective Pressure 75
3.5.2 Amino Acid Substitution Pattern: Conservative versus Radical 76
CHAPTER 4 THE EVOLUTION OF HUMAN IMMUNODEFICIENCY VIRUS TYPE II AFTER THE CROSS-SPECIES TRANSMISSION 78
4.1 THE OUTBREAK OF HIV-2 79
4.2 SEQUENCES AND ANALYSES 81
4.2.1 Sequence Data and Phylogenetic Relationships 81
4.2.2 Detection of Adaptive Evolution and Inference of Positive Selection Sites 82
4.2.3 Detection of Recombination and Inference of Recombinant Sequences 83
4.2.4 Estimation of d_N/d_S ratios (ω) for Different Lineages 83
4.2.5 Estimation of ω Assuming Lineage and Site Specific Evolution 85
4.3 RESULTS 87
4.3.1 Positive Selection Detected in HIV-2 Genes 87
4.3.2 Significant Support for Recombination in HIV-2 but not in SIVsm 93
4.3.3 Different d_N/d_S Estimates for Different Part of the Phylogeny 95
4.3.4 Positive Selection Detected in Foreground and Background Lineages 97
4.4 DISCUSSION 99
4.4.1 Evidence of Adaptive Evolution Operating in HIV-2 and SIVsm 99
4.4.2 The Impact of Recombination 100
4.4.3 Lineage Specific Variation in Selective Pressure 101
4.4.4 Evidence of Adaptive Evolution in SIVmac and SIVsm Lineages 102
CHAPTER 5 THE EVOLUTION OF A SIMIAN IMMUNODEFICIENCY VIRUS (SIVMAC) POST ZOONOSIS: THE USE OF SIVMAC INFECTION OF MACAQUES AS A NONHUMAN MODEL 104

5.1 THE EMERGENCE OF A NEW IMMUNODEFICIENCY VIRUS IN MACAQUES 105
5.2 MATERIAL AND METHODS 108
5.2.1 Data Preparation and Phylogenetic Inference 108
5.2.2 Detecting Amino Acid Sites Under Positive Selection 108
5.2.3 Detecting Possible Recombinant Regions 109
5.2.4 Detecting Lineage Specific Changes in Selective Pressure 110
5.2.5 Detecting Lineage Specific Changes in Selective Constraint at Specific Amino Acids 111
5.3 RESULTS 112
5.3.1 Adaptive Evolution Detected in SIVmac Genes 112
5.3.2 Possible Recombination Detected in SIVmac Genomes 116
5.3.3 Higher d_N/d_S Estimate for SIVmac Lineages in Comparison to SIVsm 116
5.3.4 Positive Selection Detected in SIVmac and Across the Entire Phylogeny 119
5.4 DISCUSSION 123
5.4.1 Positive Selection Detected: Power and Accuracy 123
5.4.2 Adaptive Evolution Driven by CTL Recognition 125
5.4.3 Post Zoonosis: Changes in Replacement Substitution Rates 126
5.4.4 Post Zoonosis: Divergent Changes in Selective Pressure at Specific Amino Acids 127
5.4.5 Zoonosis Model Based on SIVmac Infection of Macaques 130

CONCLUSION 131

LITERATURE CITATION 135

LIST OF TABLES

Table 2.1 Parameter estimates under five models of variable ω's among sites

Table 2.2 Likelihood ratio statistics (2 $\Delta\lambda$) for comparing models of variable ω's among sites

Table 2.3 Sites identified as evolving by positive selection under M2

Table 2.4 Acceptability of amino acid substitutions at exposed and buried sites

Table 2.5 Partition of sites within different class of immunogenic epitopes

Table 3.1 Number of sites partitioned according to structural information

Table 3.2 Parameter estimates for 117 HIV-1 sequences

Table 3.3 Likelihood ratio statistics (2 $\Delta\lambda$) for 117 sequences

Table 3.4 Sites identified as evolving by positive selection

Table 3.5 Parameter estimates under amino acid substitution models

Table 3.6 Likelihood ratio statistics (2 $\Delta\lambda$) for physiochemical properties at different sites

Table 4.1 Parameter estimates for HIV-2 genes

Table 4.2 Likelihood ratio statistics (2 $\Delta\lambda$) for site-specific analysis

Table 4.3 Sites identified as evolving by positive selection under M2

Table 4.4 Parameter estimates for SIVsm genes

Table 4.5 Likelihood ratio statistics (2 $\Delta\lambda$) for hypothesis testing

Table 4.6 Sites identified as evolving by positive selection under M2

Table 4.7 Parameter estimates and likelihood ratio test statistics (2 $\Delta\lambda$)

Table 4.8 Parameter estimates and likelihood score under Branch-site models

Table 4.9 Likelihood ratio test statistics

Table 5.1 Parameter estimates under codon substitution models of variable selective pressure among sites

Table 5.2 Likelihood ratio statistics ($2\ \Delta\lambda$) for hypothesis testing

Table 5.3 Sites identified as evolving by positive selection under M2

Table 5.4 Parameter estimates under models allowing lineage specific evolution

Table 5.5 Likelihood statistics ($2\ \Delta\lambda$) for models of variable ω across the phylogeny

Table 5.6 Parameter estimates under branch-site models

Table 5.7 Likelihood statistics ($2\ \Delta\lambda$) for branch-site models

Table 5.8 Number of positive selection sites identified by M3, Model A, B and D.

LIST OF FIGURES

Figure 1.1 – The basic genomic organisation of primate lentiviruses

Figure 1.2 – The life cycle of human immunodeficiency virus

Figure 1.3 – Phylogenetic relationship of all primate lentiviruses

Figure 1.4 – The phylogenetic relationships of SIVsm/SIVmac and HIV-2 groups (A-F):

Figure 1.5 – Phylogenetic relationships of HIV-1/SIVcpz representatives

Figure 2.1 – Relative frequencies

Figure 2.2 – Physiochemical properties (acceptability) of buried sites

Figure 4.1 – Tree topology used in model R2a

Figure 4.2 – Tree topology used in model R2b

ACKNOWLEDGEMENTS

I would like to thank my supervisor Professor Ziheng Yang for his guidance, technical support, flexibility and infinite patience. I am also very grateful to Joe Bielawski for many valuable suggestions, encouragements and support. It is the effort of these two mentors that resulted in the following papers:

Yang, W., J. P. Bielawski, and Z. Yang. 2003. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J. Mol. Evol. 57: 212-221

Yang, W. J. P. Bielawski, and Z. Yang 2003. The evolution of HIV-2 post zoonosis. Manuscript in preparation.

Yang, W. J. P. Bielawski, and Z. Yang 2003. Change in selective pressure post the cross-species transmission of SIVmac. Manuscript in preparation.

Special thanks goes to Maria Anisimova for many useful discussions on the accuracy of methods and the influences of recombination. I am also very thankful to all my friends, especially Lee Summerfield and David Turner for their emotional support; without encouragements and beliefs this work would not be possible. Finally, I would like to dedicate this thesis to my mother as an appreciation for all the sacrifices she has made and her unconditional love.

INTRODUCTION

In the past two decades, we have seen many developments in vaccine, antibody and antibiotic designs. These medical and scientific progresses gave us an overoptimistic view in eradicating most pathogens that have plagued Western society. It was commonly believed that through the application of vaccines, antibodies and antibiotics, we could control most infectious diseases. Twenty years later, this view has changed, as the reality is a far cry from the pathogen-free haven we had hoped for. Even for infections successfully eradicated from the developed nations, they continued to cause health problems in developing countries. In evolutionary terms, bacteria have acquired antibiotic resistances, which become increasingly widespread. Such evolutionary advances in pathogens can lead to global health concerns, which the current treatments may struggle to deal with. To complicate the matter, many changes in our societies, in particular the population expansion and the increased mobilisation have facilitated the spread of emerging infections. It was in this era of advancements and progress that AIDS (or acquired immunodeficiency syndrome) was first discovered. The overwhelming desire to overcome this lethal infection symbolised mankind's ongoing battle with deadly pathogens.

In 1981, a sudden increase in opportunistic infections, such as pneumonia was diagnosed in relatively healthy men. It was obvious that these opportunistic infections reflected a deficiency of the immune system, which was not unlike immune suppression experienced by transplant patients. However, upon close scrutiny, it became apparent that these individuals were unable to fend off many pathogens we were commonly exposed to (Holmes 2001). Also the fatality rate was highly effectual, in that 100% mortality was observed for all cases. AIDS was first suspected to be a sexually transmitted infection, after a

link was established between prevalence and homosexuality. This sensational revelation quickly captured the public imagination and coupled with ignorance, AIDS was branded a "homosexual disease". However, many more cases were reported shortly after, but from blood transfusion recipients, haemophiliacs, drug users and heterosexual individuals. At this point, it became increasingly clear that the causative agent was infectious and can be transmitted via large volume of body fluids. The isolation of this agent in 1983 was a giant leap towards understanding AIDS (Barre-Sinoussi et al. 1983). AIDS was caused by a viral agent resembling that of human T-cell leukaemia viruses (HLTV). Subsequent epidemiological studies showed a compelling genetic similarity between AIDS and retroviruses that infected farm animals. It was then that the AIDS virus was renamed human immunodeficiency virus and was grouped with the lentiviruses. Shortly followed the isolation of HIV came the realisation that there are at least two circulating types. The first type, HIV-1 was the cause of the global pandemic, characterised by a relatively quick progression to immune deficiency. The second type, HIV-2 seemed to be slow spreading and was mainly prevalent in West Africa. Clinically it was comparatively less pathogenic, characterised by slow progression to symptoms (Hahn et al. 2000).

Once the causative agent was pronounced to be viral, its origin became the most important question. The consensus seemed to be that HIV originated from our closest primate relative, chimpanzee. However, two competing hypotheses surrounded the mode of this transmission. One theory proposed that this transfer is accidental and facilitated by the exploitation of primates in the bush meat trade. The other hypothesis is named the "oral polio vaccine" (or OPV) theory. It was suggested that the introduction of AIDS viruses was the result of contaminated polio vaccine prepared using chimpanzee kidneys (Sharp et al. 2001). This theory points the origin of HIV-1 M group (the group responsible for the

pandemic) to the administration of polio vaccine in Belgian Congo during the 1950s. The introduction of HIV-1 O, N groups and HIV-2 were thought to be the results of similar vaccine trails in west central and West Africa (Hahn et al. 2000). The research group responsible had an access to a primate centre, where chimpanzees were used to test the safety of polio vaccine and other research projects. Chimpanzee and sooty mangabey kidneys were allegedly harvested for production of polio vaccine. The kidneys of infected monkeys were thought to have contaminated the batch of OPVs and hence facilitated the transfer of AIDS viruses. Although there is no direct evidence supporting this theory, once again this sensational suggestion drawn from circumstantial evidence received enough attention for it to be examined closely (Sharp et al. 2001). Several lines of independent evidences opposed this OPV theory. First, it is almost certain that the preparation techniques would have led to the loss of SIV viability. Second, the experimental analyses of remaining vaccine sample did not suggest the use of monkey kidneys (Hahn et al. 2000). Third, the primates used for vaccine safety testing were species *P.t.schweinfurthii* and *P.paniscus*, whereas HIV-1 is most closely related to SIVcpz isolated from *P.t.troglodytes*. Fourth, the origin of M-group is estimated to be at least 10 years prior the vaccine trial (Hahn et al. 2000). Finally, under the OPV model, the current phylogenetic relationship of HIV-1should not be observed. The divergence of HIV in chimps followed by independent transfers would lead to the loss of subtype distinctions, given frequent recombination in the viral population. Hence, it is very unlikely that the OPV theory really provided a sensible explanation to the origin of human AIDS virus.

Since its first clinical classification in 1983, the AIDS epidemics have swept through the globe. It has resulted in over 16 million deaths world wide, with an additional 34 million people estimated to be carriers. This number is growing shockingly; the rate of new

infections was estimated to be roughly 5.9 million per year (**www.unaids.org**). To further complicate the issue was the realisation that the global epidemic appeared to be shifting. The epicentre was west equatorial Africa, but since has shifted towards countries such as Botswana and Zimbabwe, where the prevalence in adult population could be as high as 25%. Although Africa remained to be the centre of this pandemic, (with approximately 70% of all detected cases located in sub-Saharan Africa), increasingly higher rates of infections are reported for South and Southeast Asia including India and China. These densely populated countries are under increasing threats, as the spread of this deadly infection is facilitated by overcrowding. The epidemic appeared to be less severe in Western Europe, with approximately 500,000 reported cases. AIDS treatments are readily provided in the developed nations, resulting in a longer life span for the patient and slower progression to disease. However, due to these reasons there is growing concerns regarding reckless behaviours, which can lead to an increase in the spread of this infection (**www.unaids.org**). Hence, the prospect of eradicating AIDS becomes more daunting than ever. The scientific community was yet to give up this battle, sheer volumes of AIDS related researches have been generated since its first appearance in 1981. It is essential that we understand what makes this virus lethal. What governs the change of pathogenesis? What makes it so readily adaptable to different immune systems, and finally what are the steps we must take to gain better control of this epidemic? It is with these questions in mind that I began my studies, hoping to provide a small piece of the puzzle that may go towards completing the big picture.

Chapter 1 THE EVOLUTION AND ORIGIN OF PRIMATE LENTIVIRUS

1.1 THE GENOMIC ORGANISATION AND INFECTION MECHANISM OF PRIMATE LENTIVIRUSES

1.1.1 The Genetics and the Morphology of Primate Lentivirus

Lentiviruses are members of a large group of viruses known as the retrovirus. Retroviruses can be classified into seven genera and several subfamilies based on their nucleic acid sequences and genomic organisation (Foley 2000). In general, lentiviruses contain a positive-stranded duplex RNA genome, which is associated with nucleocapsid proteins (NC). This duplex genome is contained within a hexagonal core that is primarily made of capsid proteins (CA). The matrix proteins (MA) associate the capsid core with the viral envelope structure. The viral envelope structure consists of glycoprotein (Gp160) in complex with a lipid bilayer. Gp160 tends to form trimers that resemble spikes, which protrude out of the lipid bilayer. The glycoprotein Gp160 can be broken down into two subunits, the large subunit (Gp120) and the small subunit (Gp41). The large subunit interacts with CD4 and chemokine receptors and the small subunit facilitates viral/host membrane fusion (Novembre 2001).

The typical, (probably ancestral) structure of primate lentiviral genome is comprised of three structural genes (*gag*, *pol* and *env*), five accessory genes (*vif*, *vpr*, *tat*, *rev* and *nef*) and two long terminal repeats (LTRs). The genome is approximately 10kb in length and is highly compact, with many overlapping regions. Each gene at least overlaps in part with one other reading frame. The reading frames of *tat* and *rev* have two exons, with exon two located within the coding region of *env* (see Figure 1.1). Many primate lentiviruses, such as SIVsm, SIVmac, HIV-2, SIVcpz and HIV-1 contain one additional accessory gene, *vpx* or *vpu* (*vpx* in SIVsm, SIVmac and HIV-2 and *vpu* in SIVcpz and HIV-1).

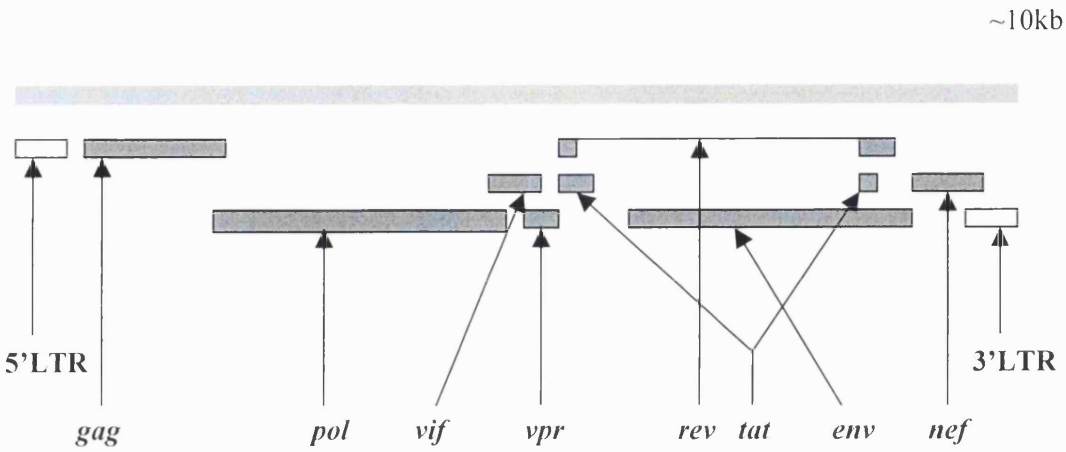


Figure 1.1 – The basic genomic organisation of primate lentiviruses: The typical length of the genome is approximately 10kb, with ▭ representing the full-length genome, ▭ representing the complete coding regions (CDS) and ▭ representing the non-coding sections.

The precise functions of these accessory proteins are yet to be fully understood. Substantial sequence similarities are observed between *vpx* and *vpr* and they are thought to inhibit cell growth and promote cell cycle arrests. However, the inhibitory effects of *vpx* appeared to be species and cell dependent (Chang et al. 2000). The *vpu* gene is unique to SIVcpz (the SIV from chimpanzee) and HIV-1. The product Vpu is associated with CD4 receptor degradation, virion release enhancement, and down-regulation of MHC-1 complex (Fischer and Sansom. 2002). The other four accessory proteins, Vif, Tat, Rev, and Nef together with Vpr are common to all primate lentiviruses. Tat and Rev are regulatory proteins involved in modulating the transcription and translation of the viral genome. The function of Vif is less well understood, though it is thought to enhance the infectivity of the virus, by assisting viral

assembly. Nef protein is implicated with the intensity of virulence, as deletion of the gene results in a less pathogenic virus. However, the precise involvement of Nef is unclear, though its expression is associated with downregulation of CD4 and MHC complexes. Thus, all the accessory proteins work in cooperation with the viral enzymes to produce infectious virion that are packaged and released into the surrounding in a process called budding (Novembre 2001).

1.1.2 The Life Cycle of Primate Lentiviruses

The production of viable progeny relies heavily on the proper function of all viral proteins, as even partial deletions can result in a defective virus. Hence, the life cycle of primate lentivirus appears to be well refined and highly regulated. The cycle begins at the invasion of T-helpers (CD4+). Although lentiviruses are capable of replication in other cell types, CD4+ cells remain to be the primary target. The binding of Gp120 (a product of *env*) to CD4 leads to the exposure of the secondary receptor. The secondary receptor belongs to a large family of receptors called chemokine receptors. The uses of different chemokine receptors give rise to viral tropism (Horuk 2001). Upon the binding of secondary receptor, Gp120 undergoes a series of conformational changes, which promotes the fusion of viral/host membranes (Dong et al. 2001). The fusions of these membranes are mediated by Gp41. As a result of this fusion, the virus core enters the host cytoplasm. Once inside, the core uncoats and undergoes rearrangements to produce a pre-integration complex (Bukrinsky 2001). This is characterised by the loss of CA and the attraction of MA (a product of *gag*). This rearrangement produces a high molecular weighted nucleoprotein complex, which regulates the early stages of replication, including the production of pro-viral DNA. Reverse transcription of RNA into

DNA takes place in the viral core, as the enzyme reversetranscriptase (produced by *pol*) is also packaged into the viral genome (Novembre 2001). The RNA genome is transcribed into two strands of DNA, the strand complementary to the RNA is called the minus strand and the other one, the plus strand. Nucleocapsid protein (NC) (produced by *gag*) is thought to be involved in the disassociation and the reassociation of strands. This core complex is transported into the host nucleus, where the viral core is degraded and the viral DNA is prepared for integration (Pedersen and Duch. 2001). Integration of the pro-viral DNA is mediated by integrase (produced by *pol*), which cleaves and ligates the host DNA. The substrates for this insertion are long terminal repeats (LTRs) and the insertion is not location specific in regards to host DNA (Pedersen and Duch. 2001).

The transcription and translation of the provirus can be divided into two phases, the early and the late phase. The early phase is marked by the expression of mRNAs encoding viral regulatory proteins, Tat, Rev, and Nef, which act as a signal for the mass production of structural proteins that occurs during the late phase. The LTRs also acts as the promoter and the enhancer for transcription (Novembre 2001). Regulatory protein Tat and Rev are thought to play a role in the transcription and the exportation of incompletely spliced viral mRNA. During the late phase, the structural proteins and enzymes are produced as uncleaved precursors. The accumulation of these Gag-Pol polyproteins signals for viral assembly and packaging. Assembly of primate lentivirus tends to occur at the cell membrane, where the polyprotein precursors are cleaved into mature proteins by viral protease (also produced by *pol*) (Bukrinsky 2001). Cellular machineries are exploited for the production of Gp160, which is cleaved within the Golgi apparatus. The final product is presented on the plasma membrane and incorporated into the virion envelope. The mature virion then "buds" off the host cell carrying other cellular membrane proteins, obtained from its host.

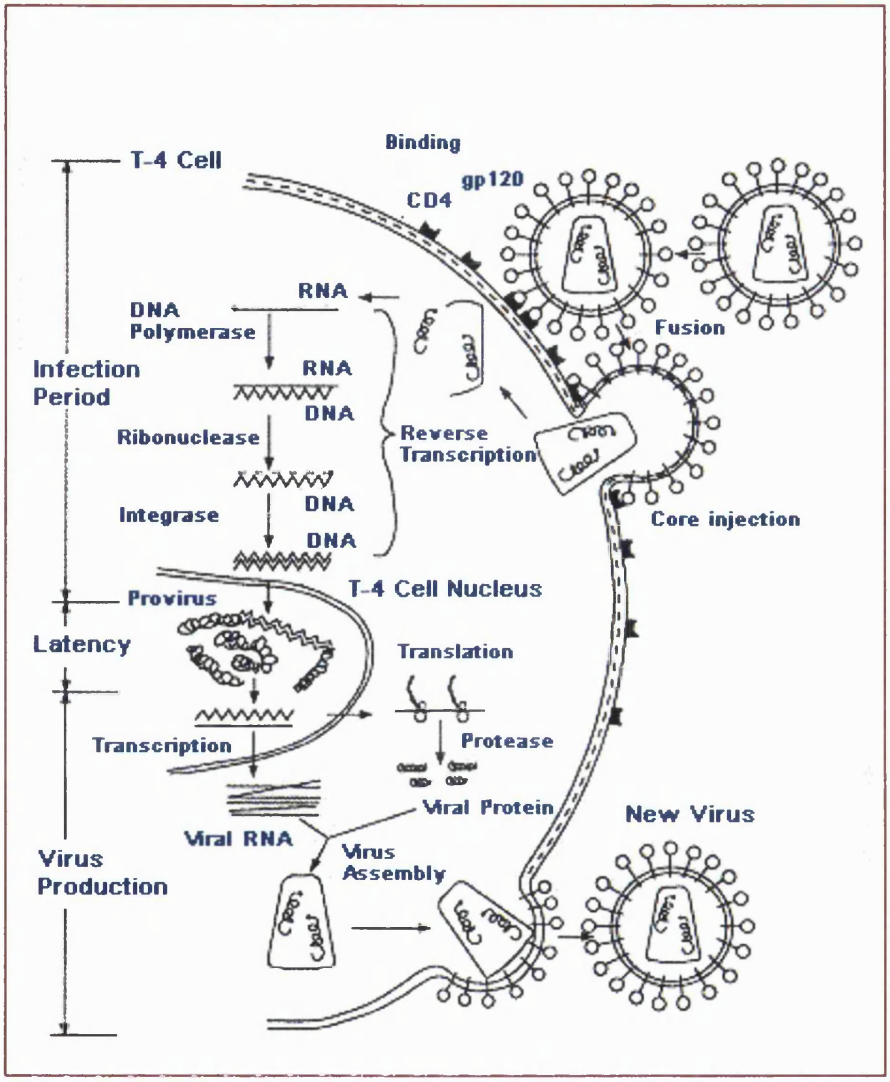


Figure 1.2 – The life cycle of human immunodeficiency virus: The life cycle of HIV-1 is highly similar to that of other primate lentiviruses. The infection starts at binding of Gp120 to CD4 and fusion of the membranes. The injected core then uncoats and reverse transcribes into pro-viral DNA. The viral DNA inserts into the host chromosome and initiates transcription. The viral mRNA is exported into the cytoplasm for the production of structural polyproteins. The congregation of these precursors starts viral assembly and budding. The

cleavage of Gag-Pol precursors into mature proteins and the formation of viral envelope complete the life cycle. **http://biosci.usc.edu/documents/gifs/fig4.3.gif**

1.1.3 The Virulence

The virulence of primate lentivirus ranges from lethal to avirulent. The primates that are natural hosts to these viruses are often without disease. However, these viruses are capable of inducing immune deficiency related disorders when introduced to foreign hosts (cross-species transmission). The severity of the acquired infection is characterised by progression to general immune deficiency, which can differ dramatically. Some cross-species transmission does not appear to be disease inducing, as observed in the SIVmnd-1 infection of mandrills. The same progenitor (SIVsm) gives rise to two different viruses (HIV-2 and SIVmac) with contrasting pathogenicity. Long–term progression to AIDS is observed for most HIV-2 infections, whereas infections in macaque are highly virulent. In extremity, the virus causes fatality within days post transmission.

Little is known regarding the factors determining the variation in virulence. However, the rate of viral replication and the intensity of host immune response appear to influence pathogenicity (Rey-Cuille et al. 1998). The rate of viral replication reflects the viral load within the host, which is the quantity of the virus, persisted in the system after viral clearance. A high viral load at the asymptotic stage tends to increase the risk of disease progression, as observed in SIVmac and HIV-1 infections. Hence, a high viral load of 10^4 viral DNA replicates/10^6 host cells leads to fast disease development in human and macaques (Rey-Cuille et al. 1998). Interestingly, the non-pathogenic strains are capable of replicating at the same rate as their pathogenic relatives. Infected sooty mangabey can have a viral load

as high as observed in HIV infections and maintains a high replication rate during the course of infection without inducing disease (Holmes 2001). Many infected African green monkeys (AGM) can also tolerate high levels of viral load during the chronic phase of infection. Thus, the rate of viral replication only appears to influence pathogenicity in foreign hosts. Viral load and viral clearance are directly affected by the intensity of host immune response. In vitro, CD8+ cells from macaque, mangabey and AGM can inhibit viral replications. A strong CD8+ cell mediated viral clearance is important in controlling disease progression in SIVmac and HIV-1 infections, whereas a relatively weak CTL response is detected in sooty mangabey. It appears to be that the intensity of host immune response required for effective viral control differs between natural and foreign hosts (Rey-Cuille et al. 1998). Not all cross-species transmissions resulted in an increase of virulence and effective viral control is observed in naturally infected mandrills. Mandrills infected with SIVmnd-1 are thought to have acquired the infection from sun-tail monkeys. SIVmnd-1 is capable of rapid replication at the onset of infection, which does not seem to increase the risk of disease progression. Instead, a transient viral load is observed in infected mandrills, characterised by the peaking and steady decline of viraemia within 60 days post infection (Onanga et al. 2002). The CTL mediated response is also transient. The number of CD8+ lymphocytes in mandrill peripheral blood returns to normal after a rapid increase during primary infection. Hence, mandrills appear to modulate the level of viraemia by a transient immune response. Also, it appears to be that African primates have established a delicate equilibrium with these non-pathogenic viruses. Disturbances to this equilibrium can lead to changes in virulence.

1.2 THE EVOLUTION OF NON-HUMAN PRIMATE LENTIVIRUSES

1.2.1 The Origin and the Diversity of Simian Immunodeficiency Viruses

Lentiviruses can be further divided into five subgenera according to their hosts, bovine, equine, feline, ovine, and primate lentiviruses. The genetic distances between the non-primate viruses of the same host species are greater than those observed between the primate viruses. Hence it is likely that primate viruses have a non-primate origin (Foley 2000). Also the primate lentiviruses form a monophyletic cluster within the lentiviral phylogeny suggesting a single transmission event from a non-primate followed by diversification in other primate species as the infection spreads. Primate lentiviruses could be further divided into human and simian immunodeficiency viruses (HIVs and SIVs). No natural lentiviral infections are detected in new world and Asian primates. To date, over 20 species of African primates are infected with SIVs. Most natural hosts appear to harbour their own monophyletic lineage of lentivirus (Beer et al. 1999). The viruses are genetically distinct and are designated according to their host names.

These SIV infections form six distinct clades consist of a) SIVcpz from chimpanzees, b) SIVsm from sooty mangabey, c) SIVagm from African green monkeys, d) SIVsyk from Sykes' monkeys, e) SIVlhoest from L'Hoest monkeys, and f) SIVcol from colobus moneys. Substantial sequence divergence was observed among the clades, with up to 50% of amino acid differences in the *pol* gene (Courgnaud et al. 2001). Difference in host subspecies contributes to the extensive sequence variations observed within the three major clades (SIVcpz, SIVagm and SIVlhoest). The currently documented SIVcpz strains were isolated from two subspecies of chimpanzees, *Pan troglodytes troglodytes* and *Pan troglodytes schweinfurthii*. The sequence diversity (of SIVcpz) between the two subspecies is greater

than within the same species. In *pol*, 24% of amino acid differences were observed between the subspecies and 9-13% within the species, indicating possible host-dependent evolution (Gao et al. 1999). Co-evolution was also observed in SIVagm, which is divided into four distinct subtypes according to the host subspecies. SIVlhoest clade is comprised of SIVlhoest, SIVsun and SIVmnd-1 lineages. As their hosts, SIVlhoest and SIVsun are the closest relatives to each other, with host-dependent evolution promoting their sequence divergence (Takehisa et al. 2001). The phylogenetic relationship of all primate lentiviruses (without the recombinant viruses) is as shown in figure 1.3.

The infected primates that do not carry their own lineage of lentivirus were thought to have acquired the infection via cross-species transmission (zoonosis). Cross-species transmissions are rare in comparison to within species transmissions. Many primates share overlapping habitats, yet each species harbours their own separate lineage of SIV. Most notable is that African green monkey (AGM) and chimpanzee share the same habitat yet there is no observed zoonosis between the two. It appears to be that host species behaviours and species restriction of the viruses are important factors in determining the success of zoonosis (Charleston and Robertson. 2002). The lentiviruses isolated from baboons, mandrills and macaques are such examples. Baboons and the vervet subspecies of the AGM share the same habitat and several successful cross-species transmissions are observed in the wild. Mandrills are also found to be naturally infected with two different types of SIV. SIVmnd-1 is isolated from mandrills in southern and central Gabon and it is a close relative of SIVlhoest. In contrast, SIVmnd-2 is isolated from north Gabon and Cameroon and it shows more similarity to SIVdrl. It is probable that two separate zoonotic events from different hosts gave rise to these infections (Souquiere et al. 2001). Successful transmissions of SIVsm from sooty mangabey to rhesus macaque were observed only in captivity, but not

in wild, as macaques are naturally occurring outside Africa (Rensburg et al. 1998).

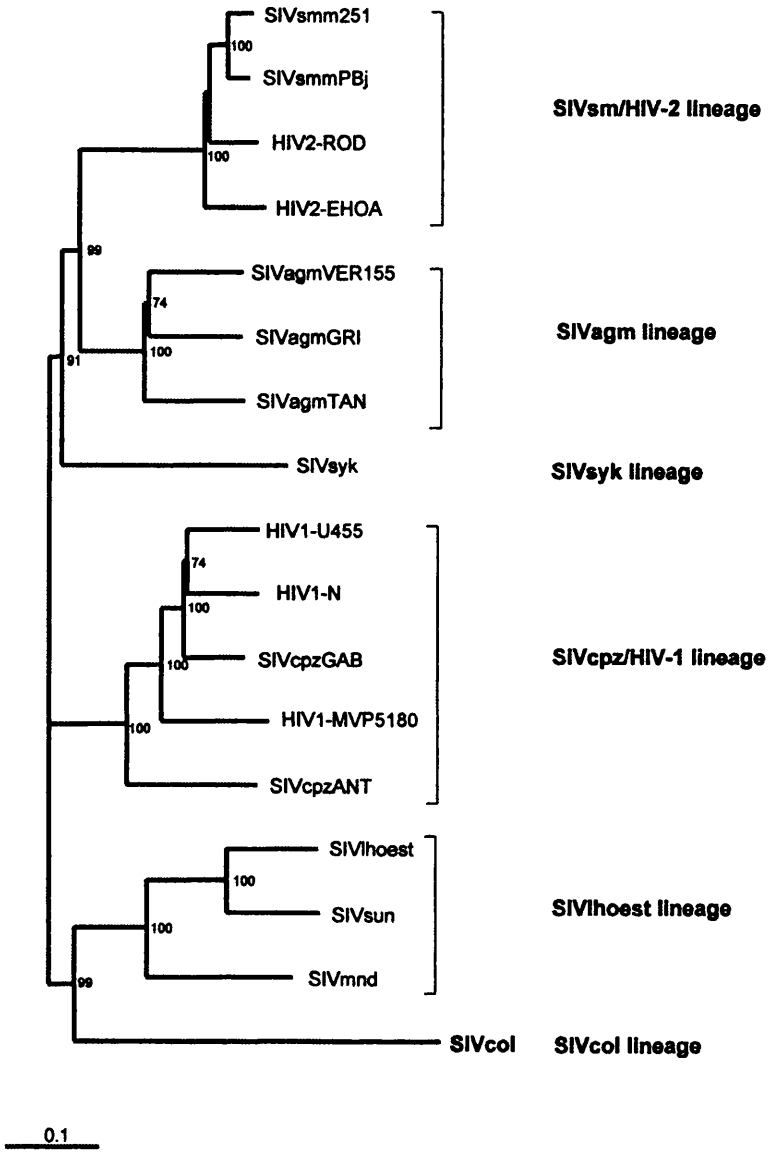


Figure 1.3 – Phylogenetic relationship of all primate lentiviruses: This relationship is inferred by neighbour joining using a concatenated dataset (Gag-Pol-Vif-Env-Nef). The branch lengths are scaled as 0.1 amino acid replacements per site. A bootstrap of 1000 replicates is used to estimate the support values. Only bootstrap support values exceeding

75% are shown. *Adapted from Courgnaud et al. 2001.*

1.2.2 Selection and Recombination

The sequence diversity of primate lentiviruses is influenced by the viral replication rate, adaptive evolution, recombination, and coevolution. As previously discussed, primate lentiviruses are capable of replicating to a high number within a single infected host. This replication process is error-prone, due to the inability of viral reverse transcriptase to correct misincorporated nucleotides. Hence, the sequence diversity is partially influenced by the viral replication rate. In general, an error rate of 10^5 substitutions /site/replication cycle for the reverse transcriptase appears to be the maximum error rate for a genome of 10kb (Overbaugh and Bangham. 2001). Accumulation of deleterious mutations and the consequent reduction of fitness are expected if error rate exceeds the maximum. The deleterious mutations have a reduced fixation rate, as purifying (negative) selection is expected to operate throughout the genome. Hence, the mutation rate of the virus is constrained by the structural and functional requirements of proteins, which sets a limit for sequence diversity. Lentiviral enzymes and structural proteins encoded by *gag* and *pol* are especially well conserved over time. In general, approximately 50% of amino acid identity is observed in *pol* across all the SIV lineages. In contrast, the Env protein is much more variable, with up to 75% of amino acid differences observed among all the SIVs (Courgnaud et al. 2001). Insertions, deletions and amino acid replacements seem to be better tolerated in Gp160 than in other structural proteins, suggesting that different genes are subjected to different intensities of selective constraint.

The selective pressure operating at the protein level is generally measured by the replacement/silent substitution rate ratio (d_N/d_S). These rates are defined as the number of

nonsynonymous (replacement) substitutions per nonsynonymous site (d_N) and that of synonymous (silent) substitutions per synonymous site (d_S). At highly conserved sites (such as the active site of reverse transcriptase), amino acid replacements are likely to be deleterious and their fixation rate is reduced by purifying selection. Hence, $d_S > d_N$ and d_N/d_S (ω) <1. If the substitutions have no selective influence, then the nonsynonymous changes are as likely to be fixed as the synonymous ones. Thus, $d_S = d_N$ and $\omega = 1$. At highly immunogenic sites, nonsynonymous substitutions can be promoted and fixed at a faster rate than silent changes; result in $\omega > 1$ that is characteristic of adaptive evolution. Hence, the estimation of d_N/d_S (or k_A/k_S) can be used as a tool to elucidate variation in selective pressure across the viral genome.

Adaptive changes are often seen as the appearance of viral mutants that are capable of escaping host immune recognition even after a single generation. These mutants often carry subtle modifications in the immunogenic epitopes, which can represent a substantial proportion of the genome (e.g. immunogenic epitopes are mapped to most HIV-1 genes). Hence, adaptive evolution can play a dominant role in generating sequence diversity in primate lentiviruses. Many papers have documented the extensive antibody and CTL responses towards the surface glycoprotein Gp120, where an accumulation of nonsynonymous substitutions is noted as a result of adaptive evolution (Courgnaud et al. 1998; Evans et al. 1999; Overbaugh and Bangham. 2001). In pathogenic infections such as SIVmac, an elevated nonsynonymous rate (d_N) is almost always observed at Gp120 and /or CTL epitopes of *nef* (Courgnaud et al. 1998; Kaur et al. 2001; Kirchhoff et al. 1999; Shpaer and Mullins. 1993). It appears to be that Gp120 and Nef are evolving by positive selection in most pathogenic infections (Shpaer and Mullins. 1993; Zanotto et al. 1999; Yamaguchi-Kabata and Gojobori. 2000). However, in non-pathogenic infections the d_N/d_S ratios of these

proteins could vary greatly between different viruses. In the avirulent infections of SIVagm, low d_N/d_S ratios are detected in Gp120 and Nef (Muller-Trutwin et al. 1996; Shpaer and Mullins. 1993). In the equally asymptomatic infections of SIVsm, an elevated d_N/d_S ratio is observed in Gp120 and Nef during the chronic phase (Courgnaud et al. 1998; Kaur et al. 2001). Hence, the fixation and accumulation of replacement substitutions in a protein is greatly dependent on selective pressure, which in turn generates genetic diversity.

Sequence divergence is also greatly enhanced by recombination, which occurs frequently in retroviruses, largely the result of slip-strand replication mechanism. The reverse transcriptase is capable of switching between the two strands of RNA templates during pro-viral DNA synthesis. Hence, recombinants could only be generated from viruses that replicate within the same cell. If the host is coinfected (or superinfected) with one or many different groups of the same virus, then such recombination is considered as inter-group/subtype recombination (Robertson et al. 1995). Inter-group/subtype hybrids can have a mosaic genome that is often characteristic of circulating recombinants. These hybrids could be detected by phylogenetic reconstruction, where different parts of the genome tend to cluster with different groups. Using this approach, four primate lentiviruses (SIVrcm, SIVagm.sab, SIVmnd-2, and SIVdrl) are identified as the hybrids of at least two genetically distinct lineages (Onanga et al. 2002). Extensive researches have suggested that SIVagm.sab is a hybrid of SIVsm and SIVagm, with recombination occurring at the 3' of *gag* and 5' of *pol* (Beer et al. 1999). SIVmnd-2 appears to be the recombinant of SIVrcm, SIVsm and SIVmnd-1, with a highly mosaic genome. Phylogenetic analyses showed the clustering of *gag*, *pol*, *vif*, *vpx* and *tat* with SIVrcm and *vpr* with SIVsm. The *env* and *nef* appears to be more closely related to SIVmnd-1 than SIVrcm or SIVsm (Souquiere et al. 2001). It is probable that the ancestral SIVmnd-2 is a hybrid, as its most conserved genes (*gag* and *pol*)

are related with SIVrcm, which is a recombinant. SIVrcm seems to share the same evolutionary origin with SIVsm in *gag* and SIVcpz in *pol*, indicating possible recombination events early in the history of the virus (Souquiere et al. 2001). Hence, recombination between different lentiviral lineages not only result in viable hybrids but can promote substantial genetic diversity.

1.3 THE ORIGIN OF HUMAN AIDS VIRUSES

1.3.1 The Origin of HIV-2

The lentiviruses that cause acquired immunodeficiency syndrome (AIDS) in humans are designated human immunodeficiency virus types 1 and 2 (HIV-1 and HIV-2). Their closest relatives are primate lentiviruses. As previously discussed, most infected primates appear to either harbour their own monophyletic lentiviral lineage or have acquired the infection via zoonosis. Cross-species transmission or zoonosis can occur readily in free interacting hosts with overlapping habitats. Humans do not appear to carry their own lineage of lentivirus; instead two very different lineages are isolated. Extensive researches have shown that the two viruses are phylogenetically more related to other primate lentiviruses than to each other. It has long since been hypothesised that these lentiviral infections are the results of relatively recent zoonotic transmissions from simian hosts (Sharp and Li 1988; Gojobori et al. 1990).

Five criteria have been set to determine whether acquired infections can be classified as zoonotic transmissions. The viruses and their prospective progenitors are assessed by, a) resemblance in genomic organisations, b) phylogenetic relationships, c) natural occurrence in

hosts, d) habitat overlapping and e) possible routes of transmission (Hahn et al. 2000). Phylogenetic, molecular and epidemiological studies have identified SIVsm (from sooty mangabey) to be the progenitor of HIV-2 and produced convincing evidences to satisfy the above criteria. Both viral genomes codes for a unique accessory protein that is absent in all other identified primate lentiviruses. The molecular phylogeny of all lentiviruses places SIVsm to be the closest relative of HIV-2 (Foley 2000). In addition, individual strains of HIV-2 cluster with certain lineages of SIVsm, so that some subtypes of HIV-2 seem to be more closely related to SIVsm than to each other Also clustering with the HIV-2/SIVsm lineages are SIVmac lineages, which represented another independent zoonotic event from mangabey into macaque, as SIVmac infections are only observed in captivity. Thus, successful SIVsm infections can occur across the species barrier (Sharp et al. 2000). Moreover, sooty mangabeys are found to carry their own monophyletic lentiviral lineage (see Figure 1.4), which is avirulent (in its natural host), indicating its likely role as a natural reservoir. Furthermore, these naturally infected mangabeys are scattered among the West African countries, where HIV-2 is endemic. Finally, frequent human-infected mangabey contacts are made possible by the domestication of sooty mangabey and their circulation in the bush meat trade (Hahn et al. 2000).

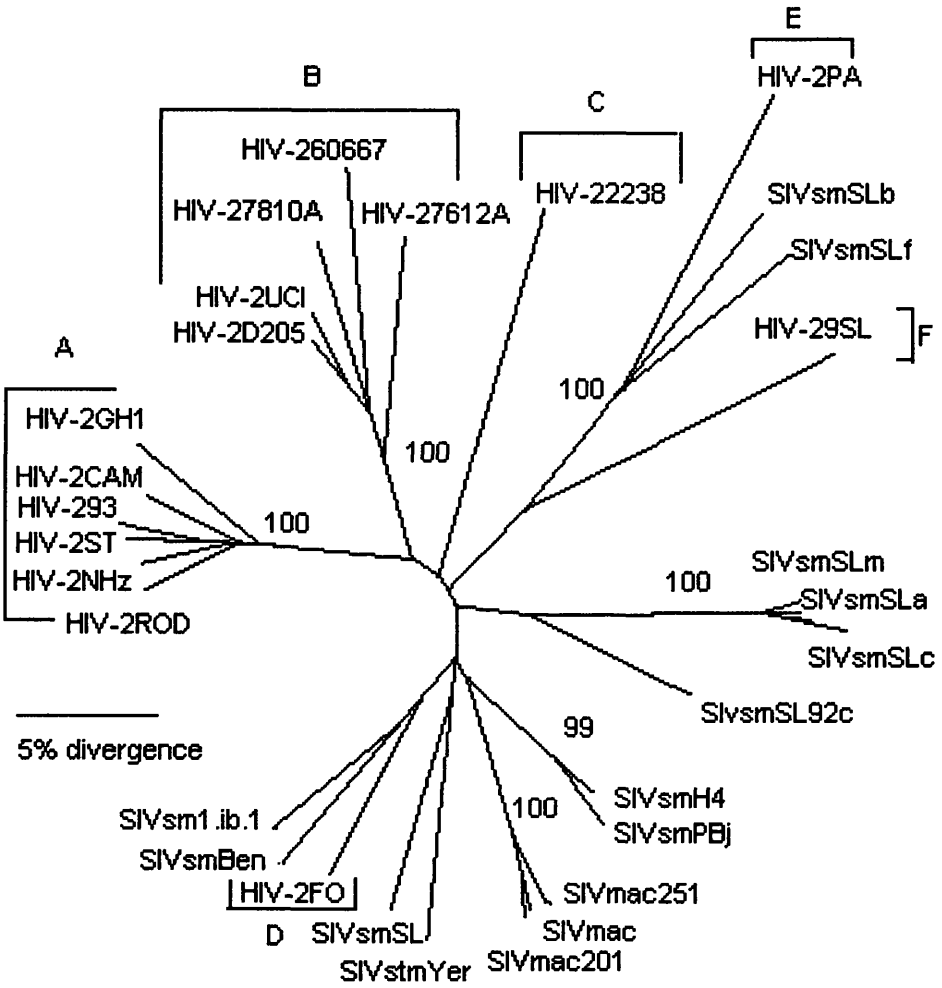


Figure 1.4 – The phylogenetic relationships of SIVsm/SIVmac and HIV-2 groups (A-F): This unrooted phylogeny is reconstructed by neighbour joining methods using a dataset of partial *gag* nucleotide sequences. A bootstrap of 1000 replicates is used to estimate the support values. Only bootstrap support greater than 50 are shown and the branch lengths are drawn to scale. *Adapted from Beer et al. 1999.*

1.3.2 The Origin of HIV-1

The origin of HIV-1 has been proven difficult to elucidate given the extensive sequence diversity observed in circulating strains. It appears to be that SIVs from chimpanzees (SIVcpz) are the closest relatives to HIV-1. The same five criteria used to determine whether SIVcpz is the progenitor of HIV-1. Some of these are readily fulfilled. Both viruses are found to be near identical in genomic organisation, with a unique accessory gene *vpu*. Frequent chimpanzee hunts and husbandry are identified as possible routes of transmission. However, the lack of SIVcpz isolates hindered the assessment of phylogenetic relatedness and indicated low host prevalence. Only three SIVcpz sequences (SIVcpzGab1, SIVcpzGab2 and SIVcpzAnt) were available for phylogenetic analyses prior year 1999 (Hahn et al. 2000). Out of the two Gabonian sequences only the full genomes of Gab1 and partial *pol* of Gab2 were characterised. Further adding to this uncertainty is the unusually large distance between the Ant isolate and the Gab1 sequence, which was placed just outside the M and N groups of HIV-1. The breakthrough came after the isolation of SIVcpzUS from a naturally infected chimpanzee of *Pan troglodytes troglodytes* subspecies (Gao et al. 1999). This new sequence was placed to be the closest phylogenetic relative of HIV-1. Moreover, the natural habitats of *P. t. troglodytes* coincide with the location where infections of HIV-1 M, N and O groups are mass circulated. Further evidence supporting this theory was produced after systematic screening of 29 captive chimpanzees from Cameroon (Corbet et al. 2000). Three more SIVcpz sequences are isolated (SIVcpzCam3, 4 and 5) representing two natural and one captive infection. The evolutionary relationship inferred using *env*, showed clustering of Cameroon sequences with the US and HIV-1 N sequences.

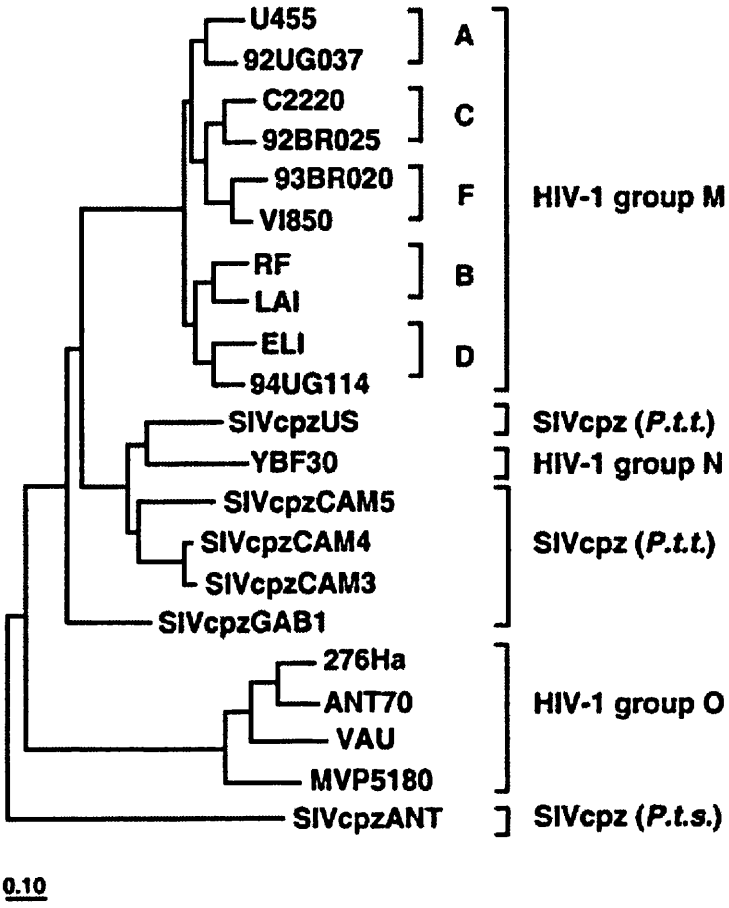


Figure 1.5 – Phylogenetic relationships of HIV-1/SIVcpz representatives: The tree is constructed by maximum likelihood using full-length Gp160 sequences. The three groups of HIV-1 (M, N and O) are labelled accordingly, though only five subtypes of group M are shown. The SIVcpz strains were classed according to chimpanzee subspecies, *P. t. troglodytes* (*P.t.t.*) or *P. t. schweinfurthii* (*P.t.s.*). Although SIVcpzCAM4 was isolated from the *P. t. vellerosus* subspecies, it is labelled as *P.t.t.* due to sequence similarity. The branch length is scaled as 0.1 amino acid replacements per site after correction for multiple hits. *Adapted from Hahn et al. 2000.*

Furthermore, the two naturally infected monkeys and the two Gabonian chimps are identified to be subspecies *P. t. troglodytes*. As expected, mitochondrial DNA analyses have shown that SIVcpzAnt is isolated from a member of the *P. t. schweinfurthii* subspecies. Hence, the assignments of chimpanzee subspecies become important in elucidating the relationships of HIV-1 and SIVcpz (Hahn et al. 2000).

The theory that SIVcpz is evolving in a host dependent manner is very attractive, as it would explain the high level of divergence observed between Ant and the other SIVcpz sequences. However, Ant is the only example of natural infection of *P. t. schweinfurthii* and no representative for *P. t. verus* is isolated to date (Sharp et al. 2001). In addition, one of the Cameroonian chimpanzees (Cam4), belonging to the subspecies *P. t. vellerosus* was found to be harbouring a virus closely related to Cam3, indicating possible transmission in captivity. Hence, given the low prevalence in the natural hosts, it is possible that both human and chimpanzee may have acquired the infection from another unidentified host. However, given the phylogenetic proximity of the viruses and their geographical relations, it is unlikely HIV-1 is originated from another source.

1.4 THE DIVERSITY AND EVOLUTION OF HUMAN AIDS VIRUSES

1.4.1 The Divergence of HIV-2 groups

As previously discussed the high replication rate and the error-prone nature of reverse transcription are partially responsible for the extensive sequence divergence observed in HIV-2 genome. To date, as many as seven genetically distinct lineages of HIV-2 are isolated

(A-G). The diversity of these clades rivalled that of M, N and O groups of HIV-1, suggesting several independent introductions into the human population. The clustering of certain HIV-2 groups with the SIVsm isolated from the same location is consistent with the notion that geographical isolation and ancestral polymorphism contributes to inter-group diversity (Holmes 2001). It is probable that some lineage is fitter than others, as seen in spread of infections. Group A and B are classed as epidemic groups and are more prevalent than nonepidemic groups C to G (Yamaguchi et al. 2000). The fitness of these groups indirectly contributes to inter-group sequence variations, as these viruses are likely to undergo evolution within hosts. As previously discussed, the extent of genetic variation is also influenced by recombination, the effect of which is most pronounced in inter-group recombination. The result of inter-group recombination is especially evident when a single host is coinfected with more than one group. Once more, the fitness of the circulating strains indirectly determines the variety of the circulating recombinants. In HIV-2, only two groups are epidemic, hence only one circulating recombinant, A/B recombinant has been reported to date (Robertson et al. 1995). To summarise, like other primate lentiviruses, geographical isolation, coevolution, ancestral polymorphism and recombination are key factors determining the high level of genetic variation observed between groups of circulating HIV-2 strains.

1.4.2 The Genetic Diversity of HIV-1

HIV-1 has been evolving at an unusually fast rate and has developed considerable genetic variation since its arrival. Phylogenetic analyses have shown that the majority of the strains fall into three principle groups M, N and O. These groups are thought to be the result of three

independent zoonotic events, indicating sequence divergence within the natural host population (Goudsmit et al. 1999). The most prevalent group that is responsible for the current pandemic is the "Main" or M group, which can be further divided into subtypes. A subtype is defined as the distinct cluster of strains from an inferred phylogeny. To date, the M group has been further divided into various roughly equidistant subtypes (A-K). The second most widespread strain is the O or the "Outlier" group, mostly located from west equatorial Africa. A third group was isolated recently in Cameroon. It is genetically distinct from the M and O group and is appropriately named "non-M, non-O", N group (Sharp et al. 2001). As previously discussed, these groups cluster with different SIVcpz lineages, indicating at least three independent introductions into human. Substantial genetic variation is observed among these groups, with the *env* sequence of group M and O differ up to 50%, indicating possible divergence of the progenitor virus (Sharp 2002). The genetic relationship of these viruses is further complicated by the rapid diversification of the M group post zoonosis. The subtypes are thought to arise from a single transmission of SIVcpz, followed by a rapid and simultaneous radiation of individual viral strains. Hence, inter-subtype variation greatly exceeds intra-subtype diversity indicating continual increase of genetic distance (Hahn et al. 2000). Moreover, there is evidence of increasing diversity within a subtype, with difference as much as 20% observed in *env*. In addition to this rapid rate of evolution, frequent recombination further promotes the diversification of HIV. To date, more than 10% of M group strains are hybrids of inter-subtype recombinations. It is this genetic flexibility (i.e. the ability to tolerate extensive genetic shuffling) that makes HIV one of the most diverse and prevalent of all primate lentiviruses.

Chapter 2 Extensive Adaptive Evolution Detected in the Human Immunodeficiency Virus Type I Genome

2.1 THE HOST IMMUNE SYSTEM AND HIV-1

Viral clearance is largely dependent on host immune response, which is often triggered by the presence of foreign antigens. In human, the primary defence mechanism against HIV-1 is a combination of humoral (antibody) and cytotoxic T-lymphocyte (CTL) responses. Antibody response is facilitated by B-lymphocytes upon recognition of foreign antigens via immunoglobulin receptors (Bondada and Chelvarajan 2001). CTL response is characterised by the detection and elimination of infected host cells (Gotch 2001). The activation of neutralising antibodies and the maintenance of CTL responses are modulated by T-helpers (Phillips, Harcourt and Price 2001). Many studies have demonstrated the involvement of antibody and CTL responses follow the invasion of HV-1 (reviewed in Fomsgaard 1999). The frequent targeting of viral proteins by antibodies, CTLs and T-helpers are noted throughout the course of infection. However the immune system itself is the main target of HIV-1, leading to the reduced CD4+ T-lymphocyte (T-helper) counts that is characteristic of HIV infection. When CD4+ cell counts drop to less than 200cellsμL^{-1}, the infected individual becomes very susceptible to other opportunistic infections and is considered to be an AIDS patient (Siliciano 2001).

The success of HIV-1 lies in its ability to escape detection by CTLs and antibodies. In the first three to four weeks of infection, a sharp increase of HIV-1 specific CTLs is observed. This is followed by rapid emergence of viral mutants with substantial sequence variation (Allen et al. 2000). Decline of CTL epitope specific lymphocyte is a direct result of viral escape by means of mutation. Direct sequencing of the entire viral genome eight weeks post infection showed an accumulation of amino acid altering mutations at CTL epitopes (Allen et al. 2000). Presumably, CTL epitopes are evolving by diversifying selection, as

neutral evolution alone could not account for such level of divergence in this short period of time (Lukashov and Goudsmit 1997). Furthermore, a sharp increase of antibody titre is also observed post HIV infection. Studies using animal models have demonstrated an increase of selective pressure on neutralizing antibody epitopes, which was sufficient to generate escape mutations (Langedijk et al. 1995; Igarashi et al. 1996; Calarota et al. 1996; McLain et al. 2001). Apparently, maximizing variation in the surface antigen allows HIV-1 to escape immune surveillance.

As discussed in chapter 1, HIV-1 has the same genomic organisation as most primate lentiviruses. Its highly compact genome encodes seven proteins and two polyproteins, which overlap either with each other or with long terminal repeats. For instance, the reading frame of *tat* overlaps with those of *vpr*, *rev* and *env*. The HIV-1 genome also is characterised by a high rate of evolution (with an average rate of 1.6×10^{-2} nucleotide substitutions per site per year) as compared with the substitution rates of human nuclear genes, estimated to be 1.3×10^{-9} substitutions per site per year (Eyre-Walker and Keightley 1999; Fu 2000). Adaptive evolution at immunogenic sites could contribute to this high evolutionary rate. Positive selection is likely to operate in regions of protein where a high level of structural specificity is not required (e.g., Walker and Goulder 2000). Thus, identification of sites with excess amino acid replacements could contribute to our understanding of positive selection and antigenic variation.

2.2 THE CROSS-SECTIONAL STATISTICAL STUDY

2.2.1 Dataset and Phylogenetic Inference

The complete genomes of 26 HIV-1 isolates were obtained from GenBank. The dataset was comprised of 5 subtype A, 7 subtype B and 14 subtype C non-recombinant isolates (Accession No AF069669-AF069673, AF042100-AF042106 and AF110959-AF110981) (Robertson et al. 1999). All the overlapping regions and the complete reading frames of *tat*, *rev*, and *vpu* were excluded from this analysis. The dataset was aligned using ClustalX (Thompson et al. 1997) and manually adjusted using GeneDoc (Nicholas et al. 1997). Due to alignment uncertainty, regions of large indels and the hypervariable regions of Gp160 (part of V1 and V2 loops) were also excluded. A Phylogenetic tree was estimated by using maximum likelihood (ML) under the model of Hasegawa et al. (1985) as implemented in the program PAUP* 4a7b (Swofford 2000).

2.2.2 Estimation of d_N/d_S (ω) Ratios Across the Genome

Parameter estimations were obtained under six codon substitution models (Yang et al. 2000): M0 (one-ratio), M1 (neutral), M2 (selection), M3 (discrete), M7 (beta) and M8 (beta&ω) as implemented in PAML (Yang 2000). The one-ratio model (M0) assumes all sites are evolving at the same rate (i.e. one ω ratio for all sites). The neutral model (M1) assumes sites in the sequence are either conserved with $\omega_0 = 0$ or evolving neutrally with $\omega_1 = 1$. M2 (selection) is an extension of M1 to account for positive selection with the addition of a discrete rate class (ω_2), which is estimated from data. The discrete model (M3) categorises

sites into K discrete rate classes, i.e. $\omega_0, \omega_1, \omega_2, \ldots \omega_{K-1}$. These rates and their respective proportions ($p_0, p_1, p_2, \ldots p_{K-1}$) are estimated from data. Three ω classes ($K = 3$) were used for this analysis. For model M7, the ω ratios estimated are approximated into a beta distribution $B(p, q)$, which is bounded within intervals (0,1). Thus, M7 does not account for positive selection. Like M1, M7 could be extended to account for sites under positive selection (i.e. sites with $\omega > 1$) with the addition of a discrete ω class. This more general model (M8) has a proportion of sites p_0 drawn from the beta distribution, while the remaining sites (i.e. $p_1 = 1 - p_0$) are assumed to be evolving at the same rate. Other parameters estimated by these models are: equilibrium codon frequencies (π), transition/transversion ratios (κ) and the branch length of the phylogeny (t). In this analysis π was calculated using nucleotide frequencies observed at the three codon positions. The branch lengths (t), measured as the expected number of nucleotide substitutions per codon along a branch and parameter κ were estimated using maximum likelihood.

2.2.3 Likelihood Ratio Tests and Bayesian Inference

The six models discussed above represented competing hypotheses. Sites in the sequence could either evolve uniformly (M0) or differently (M3). Hence, M0 is the null and M3 is the alternative hypothesis. These sites could also be evolving by neutral evolution (M1 and M7) or positive selection (M2 and M8). Thus, M1 and M7 are the null and M2 and M8 are the alternative hypotheses. These models are nested, i.e. the null model is a special case of the alternative model. Consider M3 with three ω classes ($K = 3$), if no sites fall into the ω_1 and ω_2 rate class (i.e. $p_1 = p_2 = 0$), then all sites must be evolving at the same rate ($p_0 = 1$). Hence,

M3 becomes M0 if the two rate classes are constrained. Likewise, the same parameter constriction ($p_1 = 0$) could reduce M8 to M7. In such cases, these ω classes cannot be estimated from the data, since no sites are evolving at these rates. Hence, the likelihoods for the null and the alternative hypotheses are likely to be the same. In general, the alternative model should fit the data better, with a much higher likelihood value. This improvement is often assessed by a likelihood ratio test (LRT), which compares twice the difference of the two likelihood values ($2\Delta\lambda$) in a χ^2 test. The null is rejected if the *P*-value exceeds the critical value for a given degree of freedom (Yang et al. 2000). In this analysis three LRTs were conducted to test for the competing hypotheses. The LRT comparing the one-ratio model (M0) with the discrete model (M3) was a test for among site rate variation. The comparisons of M1 with M2 and M7 with M8 were tests for positive selection.

Once positive selection is detected, an empirical Bayesian approach was used to assess the probability (posterior probability) that a site belonged to each ω class. Sites with posterior probability exceeding 95% of being from the class with $\omega > 1$, were identified from M2 and referred to as "positive selection sites" (Yang et al. 2000). Sites identified from the $\omega = 1$ rate class of M2 with posterior probability exceeding 95%, were referred to as "variable sites".

2.2.4 Amino Acid Acceptability, Protein 3D Structure and Epitope Mapping

Amino acid acceptability was measured at each site relative to four physiochemical properties: (i) polarity (Grantham 1974), (ii) volume (Grantham 1974), (iii) hydropathy (Kyte

and Doolittle 1982), and (iv) isoelectric point (Alff-Steinberger 1969). The means and standard deviations (SD) of polarity, volume, hydropathy, and isoelectric point were computed for amino acids at each site using DAMBE version 4.0.39 (Xia 2000). Amino acid acceptability at a site was measured for each physiochemical property as 100×(SD/mean). Amino acid acceptability is a measure of functional constraints acting on the amino acids at a site, with low acceptability indicating conservation of physiochemical properties at a site.

Tertiary structures were available from RCSB Protein Data Bank for all proteins except Vif. Buried and exposed sites were identified using the program WEBMOL (Walther 1997). I compared patterns of amino acid acceptability at positive selection sites and variable sites and also at exposed and buried residues of the proteins. CTL, antibody and T-helper epitopes were collected from the HIV Molecular Immunology Database 2000 (Korber et al. 2000). CTL epitopes were only used if CTLs recognised the naturally processed epitopes and both the optimal epitope and the restricting HLA molecule were defined; this subset of CTL epitopes is given in Brander and Walker (2000).

2.3 RESULTS

2.3.1 Positive Selection and HIV-1 Genes

Maximum likelihood (ML) estimates of parameters of the five major genes in the HIV-1 genome are presented in Table 2.1. Patterns of selective pressure were similar among all five genes. Here, the *gag* gene is used to exemplify my findings. Estimation of ω as an average across all sites and evolutionary history (M0) suggested that *gag* was evolving by purifying

selection. The average ω of 0.24 indicated that a replacement mutation had approximately one-fourth the chance of being fixed as compared to a silent mutation (Table 2.1). However, analyses using models allowing variable selective pressure suggested that these sites were subjected to different selection intensities (Table 2.1). Under M1, ML estimation suggested that 60% of the sites were conserved ($\omega_0 = 0$), while the remaining 40% evolve "neutrally" with $\omega_0 = 1$. Estimates under M2 indicated 1% of the sites are under positive selection, with $\omega_2 = 4.02$ (Table 2.1). Although assumptions under M1 and M2 are unrealistic for most genes, they tend to be conservative for the purposes of testing and identifying positively selected sites (Yang et al. 2000). Under M3, 72% of sites had an ω ratio of 0.05 (Table 2.1), 23% of sites had an ω ratio of 0.59 and 5% had ω ratio of 1.81.

Table 2.1 Parameter estimates under five models of variable ω's among sites

		Parameter estimates under different models					
Gene	L_C	M0 (one-ratio)	M1 (neutral)	M2 (selection)	M3 (discrete)	M7 (beta)	M8 (beta&ω)
gag	420	$\omega = 0.24$	$p_0 = 0.60$, $\omega_0 = 0$	$p_0 = 0.60$, $\omega_0 = 0$	$p_0 = 0.72$, $\omega_0 = 0.05$	$B(0.20, 0.59)$	$B(0.29, 1.18)$
			$p_1 = 0.40$, $\omega_1 = 1$	$p_1 = 0.38$, $\omega_1 = 1$	$p_1 = 0.23$, $\omega_1 = 0.59$		$p_0 = 0.95$
				$p_2 = 0.01$, $\omega_2 = 4.02$	**$p_2 = 0.05$, $\omega_2 = 1.81$**		**$p_1 = 0.05$, $\omega = 1.79$**
pol	907	$\omega = 0.32$	$p_0 = 0.68$, $\omega_0 = 0$	$p_0 = 0.68$, $\omega_0 = 0$	$p_0 = 0.81$, $\omega_0 = 0.05$	$B(0.24, 1.02)$	$B(0.33, 1.74)$
			$p_1 = 0.32$, $\omega_1 = 1$	$p_1 = 0.31$, $\omega_1 = 1$	$p_1 = 0.18$, $\omega_1 = 0.52$		$p_0 = 0.99$
				$p_2 = 0.01$, $\omega_2 = 7.20$	**$p_2 = 0.01$, $\omega_2 = 3.40$**		**$p_1 = 0.01$, $\omega = 3.49$**
vif	152	$\omega = 0.61$	$p_0 = 0.50$, $\omega_0 = 0$	$p_0 = 0.50$, $\omega_0 = 0$	$p_0 = 0.61$, $\omega_0 = 0.06$	$B(0.19, 0.29)$	$B(0.21, 0.31)$
			$p_1 = 0.50$, $\omega_1 = 1$	$p_1 = 0.47$, $\omega_1 = 1$	$p_1 = 0.36$, $\omega_1 = 0.95$		$p_0 = 0.97$
				$p_2 = 0.03$, $\omega_2 = 4.57$	**$p_2 = 0.03$, $\omega_2 = 3.74$**		**$p_1 = 0.03$, $\omega = 3.49$**
vpr	69	$\omega = 0.72$	$p_0 = 0.46$, $\omega_0 = 0$	$p_0 = 0.54$, $\omega_0 = 0$	$p_0 = 0.68$, $\omega_0 = 0.06$	$B(0.16, 0.26)$	$B(0.29, 0.78)$
			$p_1 = 0.54$, $\omega_1 = 1$	$p_1 = 0.37$, $\omega_1 = 1$	$p_1 = 0.23$, $\omega_1 = 0.73$		$p_0 = 0.91$
				$p_2 = 0.09$, $\omega_2 = 3.81$	**$p_2 = 0.09$, $\omega_2 = 2.73$**		**$p_1 = 0.09$, $\omega = 2.71$**
env	694	$\omega = 0.79$	$p_0 = 0.46$, $\omega_0 = 0$	$p_0 = 0.46$, $\omega_0 = 0$	$p_0 = 0.62$, $\omega_0 = 0.08$	$B(0.21, 0.36)$	$B(0.31, 0.68)$
			$p_1 = 0.54$, $\omega_1 = 1$	$p_1 = 0.43$, $\omega_1 = 1$	$p_1 = 0.27$, $\omega_1 = 0.82$		$p_0 = 0.89$
				$p_2 = 0.11$, $\omega_2 = 4.67$	**$p_2 = 0.11$, $\omega_2 = 3.30$**		**$p_1 = 0.11$, $\omega = 3.18$**
super gene	2473	$\omega = 0.34$	$p_0 = 0.74$, $\omega_0 = 0$	$p_0 = 0.74$, $\omega_0 = 0$	$p_0 = 0.87$, $\omega_0 = 0.10$	$B(0.74, 0.26)$	$B(0.27, 0.90)$
			$p_1 = 0.26$, $\omega_1 = 1$	$p_1 = 0.24$, $\omega_1 = 1$	$p_1 = 0.11$, $\omega_1 = 1.03$		$p_0 = 0.95$
				$p_2 = 0.03$, $\omega_2 = 5.88$	**$p_2 = 0.03$, $\omega_2 = 5.08$**		**$p_1 = 0.05$, $\omega = 3.58$**

L_C is the number of codons after the removal of alignment gaps. p_i's are the proportion of sites assigned to an individual ω category or to a beta distribution with parameters p and q. ω ratios greater than 1.0 and corresponding proportion of sites are indicated in bold. Super gene was constructed by concatenating all five protein coding genes.

Under M3, 5% of the sites were inferred as positive selection sites, this includes the 1% of sites identified under M2. In comparison, M2 is the more conservative model, as only sites under stronger selective pressure ($\omega = 4.02$) were inferred as positive selection sites. Under M8 the beta distribution was highly skewed towards the right, and the freely estimated ω class indicated a small proportion of sites (~ 5%) evolving by positive selection ($\omega = 1.79$). Furthermore, the likelihood ratio statistics confirmed that selective pressure varies among sites (Table 2.2). The null one-ratio model (M0) was rejected in favour of M3 as $2\Delta\ell =$ 361.11 greatly exceeded the critical value of 13.28, at $\alpha = 0.01$. Moreover, LRTs also indicated that positive selection was acting on a small subset of sites in *gag*, as the neutral null models (M1 and M7) were rejected in favour of the models that account for positive selection (M2 and M8) (Table 2.2).

Table 2.2 Likelihood ratio statistics ($2\Delta\lambda$) for comparing models of variable ω's among sites

Gene	M0 vs. M3 ($\chi^2_{0.01,\,4} = 13.28$)	M1 vs. M2 ($\chi^2_{0.01,\,2} = 9.21$)	M7 vs. M8 ($\chi^2_{0.01,\,2} = 9.21$)
gag	361.11	20.13	18.36
pol	524.86	82.54	84.29
vif	160.80	22.79	15.91
vpr	88.66	13.73	14.66
env	1666.43	550.63	382.46
super gene	1137.28	1047.95	252.42

In each gene, the majority of sites were subjected to strong functional constraints, with ω close to zero. Among the remaining sites a large fraction were evolving under weak functional constraints, with ω ratios between 0.5 and 0.95 (M3, Table 2.1). Most importantly, a small fraction of sites in each gene was evolving by positive selection, with $\omega > 1$ (Table 2.1). In each gene, LRTs indicated that some variation in selective pressure was due to positive selection (Table 2.2). Parameter estimates under M2 and M8 were consistent with M3 in indicating evolution by positive selection at a small fraction of sites (Table2. 1). A super gene sequence was constructed by concatenating all five protein-coding genes and analysed using the same approach. The results were identical to single gene analyses, indicating the method is robust to sequence length (Tables 2.1 and 2.2).

2.3.2 Amino Acid Diversity, Protein Tertiary Structure and Immunogenic Epitopes

At a threshold posterior probability of 90%, M2, M3 and M8 identified 112 positive selection sites. I also conducted separate analyses of the three HIV-1 subtypes. The LRTs for each gene in separate subtype analyses indicated significant statistical support for positive selection, consistent with the combined analysis as discussed above. The results from the combined subtype analysis also supported this notion. The positive selection sites (posterior probabilities $\geq 90\%$) inferred from the subtype-specific analyses were counted and those that were present in all three subtypes were designated consensus selection sites (CSS). For each gene, the CSS was a subset of those sites identified in the combined subtype analysis (Table

2.3). This is probably because I require the CSS to be identified in all three subtypes. As the criterion for the inference of CSS was strict, so fewer sites were identified.

Table 2.3 Sites identified as evolving by positive selection under M2

Gene	Combined dataset	Consensus selection sites from Separate Subtypes (CSS)
gag	15R, **28K**, 54S, 62G, 69Q, 79Y, 84T, **91R**, 138I, 146A, 215V, 252N, 280T, 357G	28K, 62G, 69G, 84V, **91R** 146A, 280T
pol	119L, **278D**, 317S, 328K, **362Q**, 366R, **400V**, 441T 489Q, **531T**, 532T, 590V, **623T**, 638Y, **709A**, 834S, 839A	119l, 278D, **328K**, 362Q, 623T, 709A
vif	31V, 33G, 36R, 37G, **39F**, 61D, 63R, 92K, 101E, **127H**, 132R, **167T**	33G, 36R, **39F**, 92K, 127H **167T**
vpr	28N, **37I**, 41G, 48E, 55A, 60I, 77R, **84T**	28N, **37I**, 41G, 60I, **77R**, 84T
env	62D, **85V, 87V, 92N, 130K, 132T,** 173Y, 178K, 183P **187D,** 188P 190S, 200V, 230N, 231K, **232T,** 238P, 240T, 275V, 277F **281A,** 283T, 289N 291S, **295N, 308R,** 321G, **336A, 337K,** 279D **340N, 343K, 344Q,** 345A, **346A,** 351E, 350R, **362K, 389Q, 440S, 442Q,** 446S, **460N, 461S,** 467I, 500K, 607A, **612A, 619L, 620E, 621Q, 624N, 640S,** 641L, **644S,** 815L, 817A, 832V, 833V, 836A, 851L	**85V**, 87V, **130K**, 173Y, 187D, 200V, **240T**, 232T, **281A, 291S,** 308R, **336A,** 337K, 340N, 344Q, **346A**, **350R**, 362K, **440S**, **442Q**, 446S, 460N, 612A, 640S, **644S**, 815L, 817A, 851L

The sites were numbered according to the reference sequence HXB2 (GenBank accession number: K03455). Positive selection sites were identified using the Empirical Bayes approach with posterior probability P ≥ 90%, with those at P ≥ 95% in bold. Consensus Selection Sites were positive selection sites identified in all three subtypes.

I selected sites from M2 with a posterior probability of 95% and estimated acceptability of amino acids at those sites. Acceptability was measured in terms of polarity, volume, hydropathy and isoelectric point (Table 2.4). To investigate differences between positive selection and relaxed functional constraints, the acceptability of positive selection sites were compared with that of the variable sites (ω_2 = 1 rate class, under M2, with posterior probability ≥ 95%). Amino acids at positive selection sites had higher physiochemical diversity than at variable sites.

Table 2.4 Acceptability of amino acid substitutions at exposed and buried sites

	Exposed sites	Buried sites
Polarity		
Positive selection sites [a]	16.91	8.49
Variable sites [b]	5.16	5.13
Volume		
Positive selection sites	41.34	25.38
Variable sites	14.54	12.37
Hydropathy		

Positive selection sites	44.68	32.88
Variable sites	13.22	19.80
Isoelectric point		
Positive selection sites	24.56	20.01
Variable sites	8.31	1.39

Acceptability was calculated as 100×(SD/mean).

[a] Positive selection sites are partitioned into exposed (75) and buried sites (37)

[b] Variable sites are partitioned into exposed (163) and buried sites (30)

Furthermore, the mean acceptabilities at exposed positive selection sites, in terms of polarity, volume and hydropathy were substantially higher than both buried positive selection sites and exposed variable sites (Table 2.4). Similar patterns were observed when we examined separately those positive selection sites located in individual classes of epitopes (CTL, antibody, and T-helper. The immunogenic epitopes were mapped onto the primary sequence and all the codons were partitioned into epitope and non-epitope sites (Table 2.5). Note that, the different types of epitopes overlap. Only 14 positive selection sites were found to be outside all epitope regions. Half of these sites were located near an epitope (one or two residues away from an epitope). The remaining seven sites were distal to any known epitopes, but might be involved in drug resistance. One such site was associated with resistance to protease inhibitor. Two others were proximal to a drug-binding pocket. A χ^2 test was used to determine whether there was a significant excess of positive selection sites at CTL, antibody and T-helper epitope. The expected number of positive selection sites for each epitope was calculated using the frequency of observed sites in that epitope. At CTL

and antibody epitopes the observed numbers of positive selection sites were not significant, being slightly less than expected. Alternatively, a highly significant ($p = 0.0001$) excess of sites were mapped to T-helper epitopes.

Table 2.5 Partition of sites within different class of immunogenic epitopes

	CTL	Non-CTL
Total sites	1057	1134
Positive selection sites	44	68
	Antibody	Non-Antibody
Total sites	960	1231
Positive selection sites	43	69
	T-helper	Non-T-helper
Total sites	1151	1040
Positive selection sites	79	33
	Epitope	Non-Epitope
Total sites	1679	512
Positive selection sites	98	14

2.4 DISCUSSION

2.4.1 Evidence of Adaptive Evolution in the HIV-1 Genome

Many studies on HIV-1 evolution have focused primarily on single gene analysis, with separate studies on *pol, vif, env* and *nef* indicating a role for positive selection (Zanotto et al. 1999; Yamaguchi-Kabata and Gojobori 2000; Yang et al. 2000). I expanded upon previous studies by conducting an extensive analysis of the entire HIV-1 genome. I used different datasets incorporating three of the most prevalent subtypes (A, B and C) to identify sites that were under continual selective pressure. All the LRTs detected positive selection in all five genes (*gag, pol, vif, vpr* and *env*) for all three subtypes. This study is the first cross-sectional statistical analysis of all the non-overlapping regions, and is also the first to suggest that every gene in the HIV-1 genome possess some sites evolving under positive selection.

In viral evolution, ω is commonly used as a measure of selective constraints on a protein (e.g., Crandall et al. 1999; Zanotto et al. 1999; Yamaguchi-Kabata and Gojobori 2000). Early analyses estimated ω as an average across all sites between a pair of lineages and indicated no role for positive selection in HIV-1 evolution (Seibert et al. 1995; Leigh Brown 1997; Plikat et al. 1997). Recent studies, however, showed that a subset of sites in *vif, env* and *nef* are evolving by positive Darwinian selection (Nielsen and Yang 1998; Yang et al. 2000; Zanotto et al. 1999). Traditional approaches failed to detect positive selection in *env* and *nef* because most sites in those genes were evolving under purifying selection, yielding an average d_N/d_S over all sites much less than 1 (Sharp 1997; Yang et al. 2000). My findings indicated that this pattern is characteristic of all HIV-1 genes examined; most sites are evolving under purifying selection, with positive Darwinian selection acting on only a small

set of sites. Furthermore, when ω is estimated as an average over all sites, no positive selection was indicated in any HIV-1 gene. This situation is not unique to HIV-1, as methods that average d_N/d_S over sites had low power to detect positive selection in other viruses (e.g., Gojobori et al. 1990).

The Empirical Bayes approach has become a popular method for inferring positive selection sites (Haydon et al. 2001; Swanson et al. 2001; Fares et al. 2001), however, Haydon et al. (2001) recently expressed concerns about type I error rates. Recent simulation studies suggested that the Bayes approach was reliable provided that the sample size was not too small and sequence divergence was not too low (Anisimova, Bielawski and Yang. 2002). In simulations, identification of positively selected sites was reliable for a dataset with as few as 17 taxa and 0.07 expected substitutions per branch (Anisimova, Bielawski and Yang. 2002). With approximately 0.16 expected substitutions per branch under M2, M3 and M8, these datasets appear to be well within the window of sequence divergence required for reliable identification of positively selected sites. Two additional lines of evidence support this notion. First, under all three models positive selection sites with a posterior probability > 90% were almost identical. Second, positive selection sites identified in *env* were highly consistent with those identified previously (Yamaguchi-Kabata and Gojobori. 2000; Yang 2001). Yamaguchi-Kabata and Gojobori (2000) identified 33 positive selection sites, 16 of which were located in regions excluded from this analysis due to alignment uncertainty. The remaining 17 sites were identified as under positive selection in this analysis. I also identified seven additional positive selection sites, four of which were also identified in Yang's (2001) analysis. This analysis differed from the previous two studies of *env* by only three additional sites, although this dataset included sequences from two additional subtypes of HIV-1.

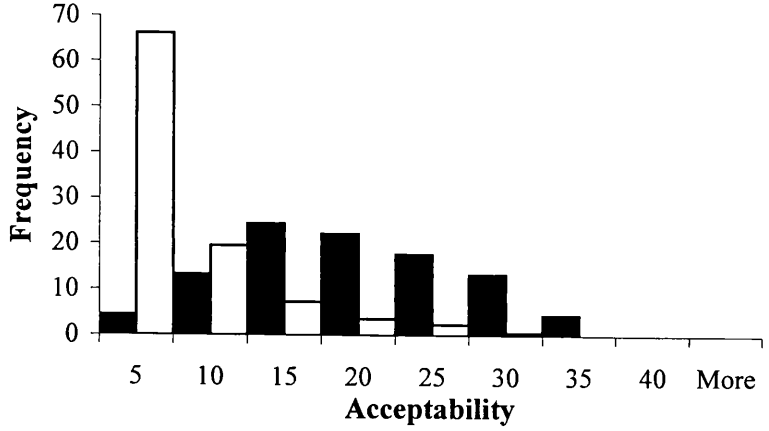
2.4.2 The Influence of Recombination

As discussed in chapter 1, the HIV-1 genome is characterised by a high recombination rate that is estimated to be as high as 3 crossovers per genome per replication event (Jetzt et al. 2000). There are two types of recombination, inter-subtype and intra-subtype. Inter-subtype recombination gives rise to circular recombinants, but does not affect this analysis because I sampled only those sequences classified to be non-recombinant at the subtype level. However, intra-subtype recombination is much more difficult to detect, limiting our ability to define a completely recombination-free dataset. With recombination, different sites may have different phylogenetic histories. Hence assuming one phylogeny may lead to false detection of sites under selection because the phylogeny is incorrect for those sites. To see the impact of phylogeny on our results, I analysed these data assuming an unrooted phylogeny. The results were highly similar to those obtained by using the estimated gene tree. Simulation studies have shown that the accuracy of Bayes inference of positively selected sites is only slightly reduced in the presence of high recombination rate (Anisimova et al. 2003). Furthermore, one expects falsely detected sites to be clustered along the primary sequence, as segments of the DNA sequence may share common history. However, in this analysis, positive selection sites were dispersed in the primary sequence and clustered on the tertiary structure. This is reminiscent of the MHC1 case, where positive selection sites were scattered in the primary sequence, but congregated on the antigen recognition site in the 3D structure (Yang and Swanson. 2002). Thus, recombination was unlikely to be responsible for sites identified to be under positive selection in this study.

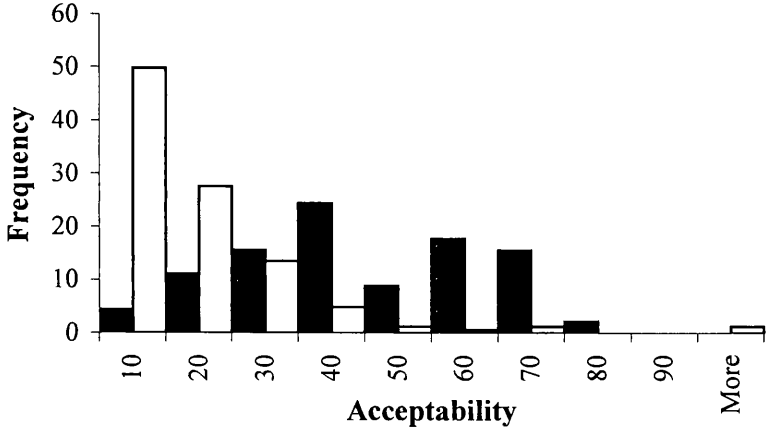
2.4.3 Amino Acid Substitution Patterns at Positive Selection Sites

There are potentially contrasting views regarding amino acid substitution pattern and positive selection. One suggests that positive selection promotes change among residues with large differences in physiochemical properties (Hughes et al. 1990; McClellan and McCracken 2001). The other suggests that positive selection must operate within certain structural constraints and amino acid substitution should be physiochemically constrained (Haydon et al. 1998; Haydon et al. 2001). My findings suggest that in HIV, substitution patterns promoted by positive selection are related to tertiary structure. Positive selection appears to promote physiochemically diverse substitutions at externally located positive selection sites (Figure 2.1), whereas less radical amino acid substitutions appear to be favoured at buried positive selection sites (Figure 2.2). Rather than competing hypotheses, these two views appear to reflect two different aspects of adaptive evolution in HIV.

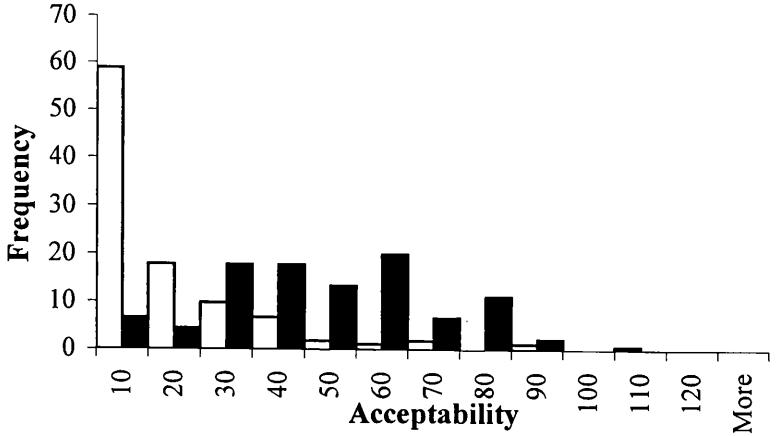
a) Polarity



b) Volume



c) Hydropathy



d) Isoelectric point

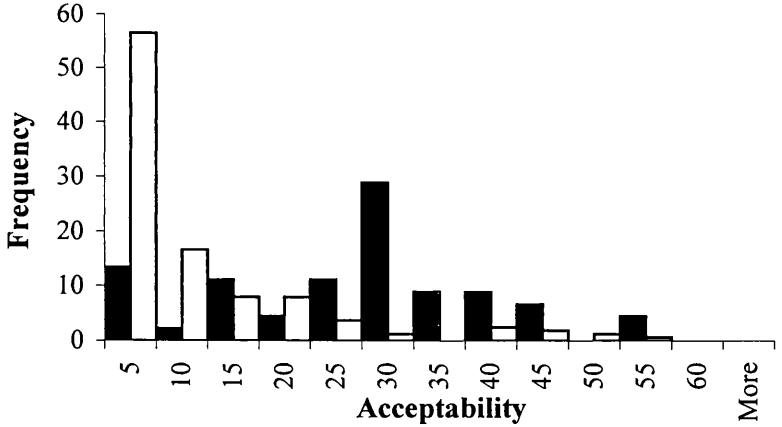


Figure 2.1- Relative frequencies: Exposed positive selection sites (black bars) and variable sites (white bars) at different amino acid acceptabilities, with the acceptability measure: a) polarity, b) volume, c) hydropathy and d) isoelectric point.

The tertiary structure of a protein could have a profound effect on all amino acid substitutions, not just those at positive selection sites. It is expected that residues in regions such as the core of the protein should be subjected to strong functional constraint, with conserved patterns of amino acid evolution. This notion was first linked to HIV-1 sequence evolution, when Yamaguchi-Kabata and Gojobori (2000) noted that endodomains of the most variable HIV-1 protein, Gp120, are more conserved than the ectodomains. I expand on this by noting that internal residues were subjected to more intense functional constraints than external residues in all HIV-1 genes examined (Figure 2.2). It is worth noting that the functional constraints on certain physiochemical properties were highly similar for both exposed and buried variable sites. In particular, properties that affect the structural specificity of the protein were subjected to strong constraint. As positive selection sites are likely to be located in regions where a high level of structural specificity is not required (Walker and Goulder 2000). Positive selection sites generally experience less constraint on physiochemical properties than variable sites. Even so, my results suggest that positive selection must operate within the functional constraints associated with internal residues. Interestingly, this is not unique to viruses, as a study of *fimA* from *Escherichia coli* yielded similar conclusions (Peek et al. 2001).

Figure 2.2 - Physiochemical properties (acceptability) of buried sites

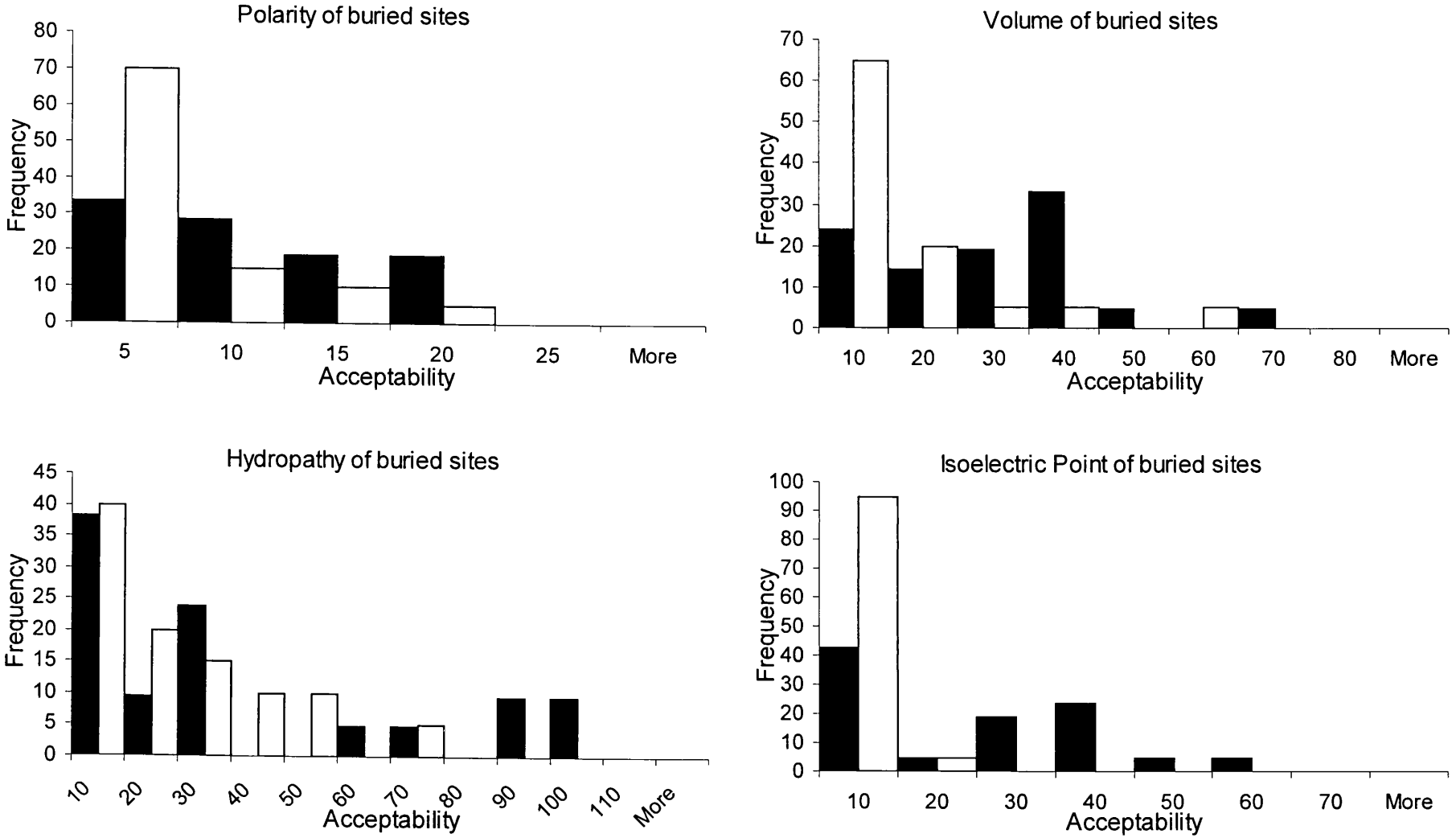


Figure 2.2- Relative frequencies: Buried positive selection sites (black bars) and variable sites (white bars) at different amino acid acceptabilities, with the acceptability measure: a) polarity, b) volume, c) hydropathy and d) isoelectric point.

2.4.4 Diversifying Selection, Antigenic Variation and Epitope Evolution

Adaptive evolution at the DNA level could either be directional or diversifying. Directional positive selection promotes a specific type of substitution that is selectively advantageous, leading to an increase of a certain phenotype, whereas diversifying selection maximizes the variation of a population. Directional positive selection is less easily detected by using the $\omega > 1$ threshold. Hence it is more likely that the sites inferred in this analysis reflect the impact of diversifying selection rather than directional positive selection. This notion is supported by the observation that the pattern of amino acid substitution is physiochemically most diverse at external positive selection sites (Hughes et al. 1990).

As the immune response is a mixture of antibody and CTL response, either antibody or CTL epitopes might experience more intense selective pressure as compared to the rest of the protein. If evolution at these epitopes strongly favours immune escape, they are expected to be evolving by positive selection for escape mutants. This notion was confirmed in experimental longitudinal studies of CTL escape, where the frequency of such escape mutations increased over time (Evans et al. 1999; Allen et al. 2000). Although 44 positive selection sites were found at known CTL epitopes, this number was slightly less than expected if the sites were sampled at random from the genome. Experimental longitudinal studies also indicate an accumulation over time of escape mutations for neutralizing antibodies (Zhang et al. 1999; Beaumont et al. 2001). Again, I found a large number of

positive selection sites at antibody epitopes (43), but this number was less than expected under a random sample of sites. It is important to note that the approach used in this study only detects sites that have been subject to recurrent positive selection over long periods of time. These results do not indicate that these epitopes are not under selection for escape mutation, rather they indicate that selection for escape mutation at either CTL or antibody epitopes has not persisted at any one site for long periods of time.

In contrast to CTL and antibody epitopes, I found a significant excess of sites under positive selection at T-helper epitopes ($p = 0.0001$). Although involved in the maintenance of antibody and CTL response, the exact role of T-helpers in viral clearance is much less well understood. It has been difficult to obtain clear evidence supporting directional T-helper escape in an experimental longitudinal study (Harcourt et al. 1998). However, Harcourt et al (1998) have isolated viral mutants that were not recognized by T-helpers, implicating involvement of T-helpers in viral control. Moreover, mounting experimental evidence indicates that T-helper identification plays a critical role in controlling HIV-1 infection (Rosenberg et al. 2000; Altfeld et al. 2001). In addition to this cross-sectional study, a statistical longitudinal study of HIV-1 has identified an excess of positive selection sites at T-helper epitopes in long-term progressors (Ross and Rodrigo. 2002) further emphasizing the importance of T-helper in viral clearance. These findings indicated that selective pressure at T-helper epitopes differs in a fundamental way from that at CTL and antibody epitopes in that it is stable at some sites over long periods of time. This finding supports the notion that T-helper epitopes should play a more important role in vaccine design (Norris et al. 2001 Altfeld et al. 2001).

From experimental studies, it is evident that escape mutations at T or antibody epitopes occur frequently as a consequence of selective pressure (Allen et al. 2000; Beaumont

et al. 2001). As many CTL, antibody and T-helper epitopes overlap, it is difficult to assess the major target of the immune system. This matter is further complicated in this study by the long divergence time of the three subtypes sampled, where the virus experience selective pressure from individuals with different HLA loci. Hence, sites that were once selected may drift once the selective pressure is off. Clearly, longitudinal studies with well-defined epitopes and known HLA loci are ideal for identifying directional escape mutations (such as Harcourt et al. 1998). However, statistical cross sectional studies have the power of identifying sites under long term recurrent selective pressure, as sites that only briefly experience selection would not have been detected. Hence, regardless of the underlying selective mechanism, sites with high posterior probabilities of evolving under positive selection identified in this study must be highly immunogenic.

Chapter 3 PATTERNS OF EVOLUTION OBSERVED IN HIV-1

3.1 Subtype-specific Variation in Selective Pressure

Previous work has suggested that subtype-specific variation in selective pressure could be observed in HIV-1 subtype B and C (Gaschen et al. 2002). This poses a problem for vaccine design, as most studies were concentrated on subtype B sequences. More importantly, subtype B is the most prevalent strain in developed nations such as America, whereas subtype C infections are mostly located in Africa. As the epicentre of HIV pandemic is in Africa, it is essential to provide a treatment that is effective in developing countries as well as in the other nations. These genetically distinct subtypes form sub-clusters according to their geographical origins (Gaschen et al. 2002). Although variations exist within these sub-clades, yet such differences are small in comparison to inter-subtype variations. As substantial sequence variations are observed across different subtypes (as reviewed by Holmes 2001), it is possible that subtype-specific patterns of evolution can also be observed across the genome. It is also likely that different subtypes may present different antigenic domains, which could result in variations in immune selection. It is therefore critical to determine the extent of subtype-specific variation in selective pressure across the genome.

3.2 Amino Acid Substitutions: Conservative Versus Radical

Many studies have been conducted to elucidate the impact of amino acid substitution on protein evolution (Hughes 1992; Hughes and Hughes 1993; Hughes 1994; Yang et al. 1998; Zhang 2000). Amino acid substitutions are classified either as conservative or radical in terms of certain physiochemical properties. Amino acid replacements within the same physiochemical group are regarded as conservative substitutions and those between groups

are termed radical. Although, it has been noted that conservative substitutions greatly exceeded radical replacement in terms of protein evolution (Zhang 2000), it is important to elucidate if this pattern is uniform across the entire protein. As discussed in chapter 2, more conservative amino acid replacements (in terms of certain property) were observed at buried residues, whereas more diverse substitutions were noted at exposed sites. Also it is equally important to determine the relationship between patterns of amino acid substitution and adaptive evolution. Several past studies have shown that positive selection can promote different types of amino acid substitution. One view indicates that positive selection drives radical changes (Hughes et al. 1990; McClellan and McCracken 2001). The other view suggests that positive selection is constrained by structural requirements and amino acid substitutions are physiochemically conservative (Haydon et al. 1998; Haydon et al. 2001). More importantly, an excess of radical nonsynonymous substitution as compared to conservative changes was used as a criterion to detect adaptive evolution, in absence of a d_N greater than the d_S (Hughes 1992; Hughes 1994). Since then other studies have shown that this criterion of detecting positive selection is not sensible, as adaptive evolution could occur without excess radical replacements (e.g. Zhang 2000). Hence, the relationship between positive Darwinian and patterns of amino acid substitution remained to be determined.

The 20 naturally occurring amino acids could be partitioned into different groups according to different physiochemical properties (Grantham 1974; Taylor 1986). Properties such as charge, polarity, volume, side chain composition, isoelectric point, and hydropathy etc, were used to partition amino acids into different groups, some of which were used to calculate distance matrices (Grantham 1974; Miyata et al 1979). It is difficult to assess how important these properties are in regards to protein structure and evolution. Also, one cannot devise a model that will account for all these properties, as it will be too parameter rich and

computationally exhaustive. Hence, amino acid substitution patterns could only be tested with a limited number of partitions at a time. In this study, I partitioned the 20 amino acids into three categories, a) charge, b) polarity and c) polarity and volume (Zhang 2000). The charge partitions separate amino acids into positive (R, H, and K), negative (D and E) and neutral (the remaining 15 amino acids). The polarity partitions organised the amino acids into polar (R, N, D, C, Q, E, G, H, K, S, T, and Y) and non-polar (the remaining eight). The polarity and volume partition is a finer division of the amino acids into six classes, a) polar and relatively small (N, D, Q, and E), b) polar and relatively large (R, H and K), c) non-polar and relatively small (I, L, M, and V), d) non-polar and relatively large (F, W, and Y), e) neutral and small (A, G, P, S, and T), and special or unique residue C. These partitions were chosen based on their close associations with the structural and functional aspects of the protein. Dense distributions of charged amino acids often reflect functional importance. For example, in the matrix protein of HIV-1 (p17), a densely negatively charged plane of amino acids is thought to be essential in membrane association (Hill et al. 1996). The polarity of amino acids is dictated by structural requirements; i.e. the ectodomain of a protein is predominantly composed of polar amino acids and the core is comprised of non-polar residues. This structural requirement is refined by another measurement, volume, which represented the size of the residue. In certain parts of the protein, for example the turn of alpha helices, the smallest non-polar residue is almost always favoured (i.e. glycine).

3.3 DATA AND MODELS

3.3.1 Sequence Data and Phylogeny Inferences

In this study, I analysed 117 sequences of HIV-1 non-recombinant genomes, sample represented all the known subtypes (A-K, from the HIV Sequence Database Kuiken et al. 2001). To see if positive selection acted on the same set of sites in strains isolated from different geographical regions, I have divided the sequences into developing (40 isolates, subtype C) and developed countries (55 isolates, subtype B). These datasets were analysed using models assuming among site substitution rate variation and amino acid substitution models. To elucidate the association between patterns of amino acid substitution and structural relationship, I have partitioned the sites into datasets composed of buried and exposed residues (Table 3.1). These datasets were then analysed using amino acid substitution models (as implemented in PAML 3.10) with conservative and radical substitutions partitioned according to the physiochemical properties defined above. To determine the influence of positive Darwinian selection on amino acid substitution pattern, I have further divided the data into two sets, candidate positive selection sites and "variable" sites (the classification of these sites were the same as described in chapter 2). These datasets were also analysed using the amino acid substitution models. As these amino acid substitution models required a topology for parameter estimation, the most likely tree was constructed using maximum likelihood, under the model of Hasegawa et al. (1985) as implemented in PAUP (Swofford 2000).

Table 3.1 Number of sites partitioned according to structural information

Protein	Buried sites	Exposed sites
Matrix (p17)	96	282
Capsid (p24)	111	519
Protease	87	168
Integrase	180	615
Reverse transcriptase	363	1293
Gp120	441	585

3.3.2 Detecting Positive Selection in HIV-1 Genome

Six codon substitution models were used to estimate parameters: M0, M1, M2, M3, M7 and M8 (Yang et al. 2000). These models use a statistical distribution to describe the substitution rate ratios (ω). The one ratio model (M0) assumes substitution rate homogeneity across all sites and evolutionary history (i.e. 100% of the sites have the same ω). Two fixed ω rate classes were assumed for M1. The sites were either conserved (with ω_0 fixed at 0) or evolving "neutrally" (ω_1 fixed at one). The fraction of sites belonging to ω_1 rate class (p_1) is estimated from the data and the remaining proportion is calculated as: $p_0 = 1 - p_1$. M2 is an extension of M1 by incorporating a third rate class (ω_2), which is estimated from data and thus allows for positive Darwinian selection. Hence, three discrete distributions were used in M2 to class the sites into, a) conserved, b) "neutral", and c) possibly positively selected. This

model is prone to multiple local optima, as the neutral model assumptions (i.e. $\omega_0 = 0$, and $\omega_1 = 1$) are highly unrealistic, often sites with $0 < \omega < 1$ is forced into the ω_2 rate class (Yang et al. 2000). Hence, two initial ω values, one less than 1, and one exceeded 1, were used to obtain the true optimum (Bielawski and Yang. 2003). M3 uses discrete distributions to approximate ω ratios into the number of classes defined by user (i.e. K = number of discrete classes). As all these ω classes are estimated from the data, this model assumes among site rate heterogeneity and can be compared to M0 in a likelihood ratio test (LRT). M7 uses a beta distribution that is bound between 0 and 1 to describe the substitution rate ration. This model is the null for M8, which is an extension of M7, by the addition of a discrete rate class ω. The alternative models (M2 and M8) could be compared to their nulls (M1 and M7) by likelihood ratio statistics ($2\ \Delta\lambda$) as tests for positive Darwinian selection. Once again, M8 is thought to have multiple optima and hence different initial values were used to obtain the true peak (Bielawski and Yang. 2003).

3.3.3 Amino Acid Substitution Models

Two amino acid substitution models were used to estimate the rate of radical and conservative nonsynonymous substitutions. The null model (i.e. the Poisson process model) is simple and unrealistic in that it assumes an equal substitution rate for any given pairs of amino acid. It also assumes a clock like evolution; i.e. a constant substitution rate among all the lineages (Yang et al. 1998). Hence, one ω ratio is estimated for all the possible amino acid substitutions. The alternative model is referred as the "general model", in that two independent substitution rate ratios were estimated from data. The amino acid substitutions

were classed into conservative and radical substitutions in terms of physiochemical partitions. The rate of radical nonsynonymous substitutions was estimated as ω_0, and the rate of conservative changes was estimated as ω_1. As the two models are nested (i.e. the null model is a special case of the general model, given that $\omega_0 = \omega_1$), they can be compared using a LRT with one degree of freedom.

3.4 Results

3.4.1 Adaptive Evolution Operating at Different Sites in Different Subtypes

Parameter estimates using 117 sequences of the HIV-1 genome (inclusive of all known non-recombinant genomic sequences of HIV-1 subtypes) have indicated a role for positive Darwinian selection across the entire evolutionary history (Table 3.2).

Table 3.2 Parameter estimates for 117 HIV-1 sequences

Gene	M0 (one-ratio)	M1 (neutral)	M2 (selection)	M3 (discrete)	M7 (beta)	M8 (beta&ω)
gag	$\omega = 0.30$	$p_0 = 0.62$, $\omega_0 = 0$	$p_0 = 0.65$, $\omega_0 = 0$	$p_0 = 0.70$, $\omega_0 = 0.07$	$B(0.26, 0.61)$	$B(0.29, 1.09)$
		$p_1 = 0.38$, $\omega_1 = 1$	$p_1 = 0.33$, $\omega_1 = 1$	$p_1 = 0.26$, $\omega_1 = 0.61$		$p_0 = 0.96$
			$p_2 = 0.02$, $\omega_2 = 3.21$	**$p_2 = 0.04$, $\omega_2 = 2.22$**		**$p_1 = 0.04$, $\omega = 2.37$**
pol	$\omega = 0.27$	$p_0 = 0.70$, $\omega_0 = 0$	$p_0 = 0.71$, $\omega_0 = 0$	$p_0 = 0.78$, $\omega_0 = 0.05$	$B(0.27, 0.99)$	$B(0.30, 1.04)$
		$p_1 = 0.30$, $\omega_1 = 1$	$p_1 = 0.27$, $\omega_1 = 1$	$p_1 = 0.20$, $\omega_1 = 0.52$		$p_0 = 0.98$
			$p_2 = 0.02$, $\omega_2 = 6.50$	**$p_2 = 0.02$, $\omega_2 = 2.98$**		**$p_1 = 0.02$, $\omega = 3.01$**
vif	$\omega = 0.53$	$p_0 = 0.54$, $\omega_0 = 0$	$p_0 = 0.56$, $\omega_0 = 0$	$p_0 = 0.64$, $\omega_0 = 0.12$	$B(0.21, 0.39)$	$B(0.21, 0.33)$
		$p_1 = 0.45$, $\omega_1 = 1$	$p_1 = 0.41$, $\omega_1 = 1$	$p_1 = 0.33$, $\omega_1 = 0.87$		$p_0 = 0.97$
			$p_2 = 0.03$, $\omega_2 = 4.49$	**$p_2 = 0.03$, $\omega_2 = 3.62$**		**$p_1 = 0.03$, $\omega = 3.12$**
env	$\omega = 0.69$	$p_0 = 0.50$, $\omega_0 = 0$	$p_0 = 0.55$, $\omega_0 = 0$	$p_0 = 0.70$, $\omega_0 = 0.10$	$B(0.19, 0.31)$	$B(0.29, 0.81)$
		$p_1 = 0.50$, $\omega_1 = 1$	$p_1 = 0.38$, $\omega_1 = 1$	$p_1 = 0.20$, $\omega_1 = 0.75$		$p_0 = 0.90$
			$p_2 = 0.07$, $\omega_2 = 4.16$	**$p_2 = 0.10$, $\omega_2 = 3.53$**		**$p_1 = 0.10$, $\omega = 3.46$**
nef	$\omega = 0.71$	$p_0 = 0.47$, $\omega_0 = 0$	$p_0 = 0.51$, $\omega_0 = 0$	$p_0 = 0.62$, $\omega_0 = 0.08$	$B(0.21, 0.36)$	$B(0.31, 0.68)$
		$p_1 = 0.53$, $\omega_1 = 1$	$p_1 = 0.39$, $\omega_1 = 1$	$p_1 = 0.27$, $\omega_1 = 0.88$		$p_0 = 0.89$
			$p_2 = 0.10$, $\omega_2 = 3.90$	**$p_2 = 0.11$, $\omega_2 = 3.17$**		**$p_1 = 0.11$, $\omega = 3.25$**
super gene	$\omega = 0.37$	$p_0 = 0.72$, $\omega_0 = 0$	$p_0 = 0.61$, $\omega_0 = 0$	$p_0 = 0.72$, $\omega_0 = 0.06$	$B(0.76, 0.24)$	$B(0.24, 0.83)$
		$p_1 = 0.28$, $\omega_1 = 1$	$p_1 = 0.32$, $\omega_1 = 1$	$p_1 = 0.21$, $\omega_1 = 0.57$		$p_0 = 0.94$
			$p_2 = 0.07$, $\omega_2 = 4.44$	**$p_2 = 0.07$, $\omega_2 = 2.23$**		**$p_1 = 0.06$, $\omega = 2.60$**

Table 3.3 Likelihood ratio statistics (2 $\Delta\lambda$) for 117 sequences

Gene	M0 vs. M3 ($\chi^2_{0.01,\,4} = 13.28$)	M1 vs. M2 ($\chi^2_{0.01,\,2} = 9.21$)	M7 vs. M8 ($\chi^2_{0.01,\,2} = 9.21$)
gag	115.21	63.28	58.77
pol	241.08	103.19	94.52
vif	180.44	47.92	24.11
vpr	100.12	29.90	27.42
env	1474.15	445.21	346.28
nef	157.70	122.56	122.43
genome	2499.05	1029.82	991.64

Table 3.4 Sites identified as evolving by positive selection

Gene	Developing country (subtype B)	Developed country (subtype C)
gag	67S, 84T, **91R**, 280T, 373S, 389I	28H, 79F, **91E**, 146P, 252N, 357S, 375T
pol	**119L**, 531T, 623P, 816L, **839T**	**119P**, 278D, 638Y, 682K, 709N, **839A**
vif	31V, 33G, 39F, 92K, 101E, 109L, **127H**, 132R, 167T	36N, **127H**
vpr	37I, 41G, 77R, 85R, 96K	60I, 84V, 91E
env	**85V**, 87V, 134L, 139N, 140T, 141N, 149M, 150E, 151K, **161I**, 169V, **190S**, 198T, 200V, 275A, 283T, 306R, 308R, 317F, 323I, 340A, 343K, **360I**, 363Q, 407N, 408T, 440S, **460N**, 462N, **464E**, 518L, 533T, 588K, 607A, **612A**, **620E**, 621N, 636N, **640S**, 641L, 644S, **674N**, 836A, 836C	84I, **85V**, 132R, **161A**, 170K, 171T, 173Y, 188S, 189S, **190Y**, 240H, 281V, 295V, 321D, 335K, 344G, 346S, 350A, **360K**, 362A, 363S, 364S, 389G, 393G, 398S, 399I, **460N**, **464D**, 465T, 500A, **612S**, **620Q**, 624D, **640N**, 648D, 667K, 648D, 667K, **674D**, 833I, 838C, 842H
nef	**8S**, 14P, 15A, **50A**, 51N, **85V**, 101I, 102H, 114I, **120Y**, 133I, 138T, 163S, 170L, **176P**, **182E**, **188R**, 194V, 198L	**8S**, 40H, 49A, **50D**, **85L**, **120Y**, 156A, **176E**, 177H, **182K**, **188Q**

The null models were rejected with statistical significance (see Table 3.3). Across the entire genome 127 positive selection sites were detected at a threshold probability of 90%. This list of sites included the 112 positive selection sites identified in chapter 2, indicating that the accuracy of my analyses were not influenced by sample size. The analyses of subtype B and C sequences indicated that positive selection, (though detected across the genomes of these subtypes) does not appear to be operating at the same set of sites (Table 3.4). Sites identified as positive selection sites in both subtypes were in bold font.

3.4.2 Conservative Amino Acid Substitutions Fixed at a Higher Rate

My findings suggested that there was a rate difference between conservative and radical amino acid substitutions at certain sites (Table 3.5). The Poisson process model was rejected with significance in these cases (Table 3.6). My results suggested that in general, amino acid substitutions tend to be physiochemically conservative in terms of polarity and volume (Table 3.5). The overall amino acid substitution rates appeared to be higher at exposed residues. However, conservative changes in terms of polarity and volume predominated at both buried and exposed sites. The difference between the rates of conservative and radical substitutions (at exposed sites) with respect to polarity was quite small, leading to difficulties in rejecting the null (Table 3.6). Interestingly, at buried residues, radical substitutions (with respect to charge) appeared to be fixed at a rate close to that of conserved changes (Table 3.5). Parameter estimates also indicated that positive selection does not always promote radical amino acid substitutions (Table 3.5). Both conservative and radical changes were fixed at almost equal rates at positive selection sites across the genome. The null model could not be

rejected for positive selection sites, as both radical and conservative changes were fixed at equal rates.

Table 3.5 Parameter estimates under amino acid substitution models

Physiochemical properties	Buried sites	Exposed sites	Positive selection sites	Variable sites
Poisson process model (one-ratio)	$\omega = 0.39$	$\omega = 0.51$	$\omega = 1.12$	$\omega = 0.59$
Charge	$\omega_0 = 0.18$	$\omega_0 = 0.22$	$\omega_0 = 0.75$	$\omega_0 = 0.16$
	$\omega_1 = 0.16$	$\omega_1 = 0.45$	$\omega_1 = 0.81$	$\omega_1 = 0.23$
	$\boldsymbol{\omega_0 / \omega_1 = 1.13}$	$\omega_0 / \omega_1 = 0.49$	$\boldsymbol{\omega_0 / \omega_1 = 0.93}$	$\omega_0 / \omega_1 = 0.69$
Polarity	$\omega_0 = 0.16$	$\omega_0 = 0.28$	$\omega_0 = 0.83$	$\omega_0 = 0.39$
	$\omega_1 = 0.24$	$\omega_1 = 0.34$	$\omega_1 = 0.99$	$\omega_1 = 0.64$
	$\omega_0 / \omega_1 = 0.67$	$\omega_0 / \omega_1 = 0.83$	$\boldsymbol{\omega_0 / \omega_1 = 0.95}$	$\omega_0 / \omega_1 = 0.61$
Polarity and volume	$\omega_0 = 0.13$	$\omega_0 = 0.24$	$\omega_0 = 0.80$	$\omega_0 = 0.17$
	$\omega_1 = 0.23$	$\omega_1 = 0.50$	$\omega_1 = 0.88$	$\omega_1 = 0.43$
	$\omega_0 / \omega_1 = 0.57$	$\omega_0 / \omega_1 = 0.49$	$\boldsymbol{\omega_0 / \omega_1 = 0.91}$	$\omega_0 / \omega_1 = 0.40$

Note: bold script shows a ω_0 / ω_1 close to 1 indicating radical substitutions were fixed at approximately the same rate as conservative changes.

Parameter estimates suggested that substitution at variable sites tends to be physiochemically conserved. The difference between the rates of conservative and radical substitutions, with respect to charge and polarity was relatively small, as compared to that of polarity and volume.

Table 3.6 Likelihood ratio statistics (2 $\Delta\lambda$) for physiochemical properties at different sites

	Buried sites	Exposed sites	Positive selection sites	Variable sites
LRT of rate difference in terms of charge				
2 $\Delta\lambda$	1.55	32.97	0.66	15.03
P-value	**0.28**	<0.0001	**0.35**	0.0001
LRT of rate difference in terms of polarity				
2 $\Delta\lambda$	8.07	**2.16**	0.28	21.99
P-value	0.004	**0. 14**	**0.59**	<0.0001
LRT of rate difference in terms of polarity and volume				
2 $\Delta\lambda$	14.54	23.67	1.35	15.20
P-value	0.0001	<0.0001	**0.24**	<0.0001

3.5 Discussion

3.5.1 Evidence of Long Term Recurrent Selective Pressure

Previous research has suggested that different sites might be positively selected in different subtypes (Gaschen et al. 2002). These authors proposed that there was a subtype-specific pattern in generating escape mutations in gp120, particularly at sites that may induce antibody responses. They estimated ω across gp120 using the same approach as discussed above. They detected subtype-specific changes in selective pressure across the protein, notably in the regions proximal and spanning V3. The V3 of subtype B is considerably more variable than that of subtype C. It appears to be that more sites at the end of V3 loop (335 – 365) are evolving by positive selection in subtype C, whereas the tip of V3 is relatively conserved with lower ω ratios. My findings supported this observation that subtype-specific selective pressure acts on different parts of the protein, presumably driving by immune system targeting at different regions. This potentially causes problems in the traditional methods of vaccine design (as discussed by Gaschen et al. 2002). In this study, I have expanded upon the previous work by noting that this subtype-specific variation in selective pressure appeared to act throughout the entire genome of HIV-1. Interestingly, some sites seemed to be evolving by positive Darwinian selection in both subtype B and C. It is possible that these sites were the targets of continual or recurrent selective pressure. To summarise, my findings are consistent with the notion that the pattern of evolution is not uniform across different subtypes. However, certain sites could be under sustained immune selection even after the divergence of subtypes.

3.5.2 Amino Acid Substitution Pattern: Conservative versus Radical

It is generally believed that amino acid substitutions tend to be physiochemically conservative (Zhang 2000, Haydon et al. 2001). Radical substitutions tend to be slightly deleterious, which is associated with a reduction of fitness (Eyre-Walker et al. 2002). Hence, radical mutations are most likely to be under purifying selection and fixed at a much lower rate than conservative changes. Only in small population size, can one expect to see a higher proportion of radical changes as compared to conserved ones (Eyre-Walker et al. 2002). In general, my findings supported this notion. Conservative substitutions predominate most of the sites in a protein. As discussed in chapter 2, certain physiochemical properties are more conserved than others. In general, substitutions that promoted the divergence of polarity and volume were fixed at a rate ½ that of conservative changes, indicating strong purifying selection. Radical changes in terms of polarity appeared to be less constrained, indicating such changes were better tolerated. Interestingly, the rate of radical substitution with respect to charge seemed to vary with different categories of sites. At buried residues, radical mutations in terms of charge were fixed at a similar rate as conservative changes. It appeared to be that the constraint of purifying selection was equally strong for both conservative and radical changes. This observation is consistent with the notion that the core residues are subjected to intense functional constraint.

Previous reports have shown that positive Darwinian selection tends to promote diverse changes in terms of physiochemical property (Hughes 1990; McClellan and McCracken 2001). Also reported were that adaptive evolution tends to favour conservative changes (Haydon et al. 2001). These potentially conflicting views were reviewed and amended in chapter 2. In this chapter, I expanded on my previous findings by noting the

general amino acid substitution trends for all positive selection sites in HIV-1. It appeared to be that conserved substitutions were fixed at a rate close to radical substitutions. As the buried positive selection sites were pooled together with exposed positive selection sites, one would expect to see the pattern described above. On the other hand, substitutions at variable sites tend to be physiochemically conservative, a trend noted and discussed in chapter 2. To summarise, my findings have shown that radical changes of certain physiochemical properties were more constricted than others. The rate of amino acid substitutions were comparatively reduced at buried residues as compared to exposed sites. Positive selection sites appeared to promote conservative and radical substitutions, whereas changes at variable sites were predominantly conservative.

Chapter 4 THE EVOLUTION OF HUMAN IMMUNODEFICIENCY VIRUS TYPE II AFTER THE CROSS-SPECIES TRANSMISSION

4.1 THE OUTBREAK OF HIV-2

In 1986, a second virus that causes immune deficiency was isolated from AIDS patients in West Africa. Its origin intrigued many, as this virus is genetically distinct from HIV-1. Yet both viruses have similar genomic organisations, infection mechanisms and clinical symptoms. The absence of these viruses in normal humans indicated possible cross-species transmission events. A close relative of HIV-2 was isolated, when a group of rhesus macaques developed immunodeficiency related disorders. The causative agent was a retrovirus designated SIVmac (Beer et al. 1999). Subsequent molecular and epidemiological studies implicated a simian origin to these "human retroviruses (HIV-1 and HIV-2)". However, SIVmac infections were not observed in wild macaques, indicating another natural host that was responsible for these transmissions. The isolation of a new SIV from a sooty mangabey in captivity (designated SIVsm) and subsequent molecular characterisations revealed it to be a close relative of HIV-2 and SIVmac. Unlike SIVmac infections, many wild and domestic sooty mangabeys in West Africa are infected with SIVsm. Hence it is likely that sooty mangabey is the natural host of these viruses. The close phylogenetic relationships between HIV-2 and SIVsm suggested that the latter is the progenitor virus, responsible for HIV-2 infections (Foley 2000; Holmes 2001)

Despite the high level of sequence divergence, the genomic organisation of HIV-2 is identical to that of SIVsm. Both viruses are characterised by the production of an accessory protein Vpx, which is absent in other primate lentiviruses. However, in its natural host, SIVsm infections do not result in immune deficiency. As is observed in macaque and human infections, the clinical symptoms only develop when the virus infects another species. A remarkable difference in pathogenicity was observed in HIV-2 and SIVmac. The progression

to AIDS was slow in humans leading to less morbidity, but could be extremely fast in macaques killing the host in a few days (Hahn et al. 2000). It appeared to be that the pathogenicity of the virus is dependent on its host. The mechanism that influences the change of pathogenicity is poorly understood. Experimental evidences have shown that changes in *nef* and the presence of N-linked glycosylation sites in Env enhance pathogenicity in HIV-2 (Esteves et al. 2001). Differences in the immune systems could contribute to this change, as different immunogenic epitopes were under selection. Hence the rate of nonsynonymous (replacement) substitution could increase after cross-species transmission, reflecting adaptation to the new host. Changes in selective pressure along a particular lineage, as well as adaptation at a molecular level could be detected by estimating nonsynonymous/synonymous substitution rate ratio (d_N/d_S).

Using this approach, Shpaer & Mullins (1993) reported a positive correlation between an increase of d_N/d_S ratios and pathogenesis for HIV-2. However, the averaging method lacks power to detect adaptive evolution at a molecular level. Also, the use of a phylogeny would have enabled the authors to detect lineage specific evolution. Hence, I expanded on previous researches using phylogenetic methods to elucidate the influence of adaptive evolution. In this study, the complete genomes of HIV-2, and SIVsm were analysed. Positive selection was detected in all major protein coding genes of HIV-2 and in one SIVsm gene. Also in SIVsm fewer sites were found evolving by positive selection. These findings suggested that SIVsm might still be immunogenic in sooty mangabeys.

4.2 SEQUENCES AND ANALYSES

4.2.1 Sequence Data and Phylogenetic Relationships

The complete genomes of 15 HIV-2 and 10 SIVsm isolates were obtained from the HIV Sequence Database (Kuiken et al. 2001). The 15 HIV-2 sequences included 12 group A and 3 group B non-recombinants (Accession No: U22047, AF082339, M31113, J04498, U38293, J04542, M30895, Z48731, D00835, M15390, J03654, L07625, X16109, U27200). The 10 SIVsm sequences comprised of the six genetic clones and four additional isolates (L09211-L09213, L03295, M83293, M80193, M80194, U72748, X14307, AF077017). All overlapping regions were excluded from the analyses that included three small genes (*tat*, *rev*, *vpx*) and the two LTRs. The datasets was aligned using Clustal X (Thompson et al. 1997) and manually adjusted using GeneDoc (Nicholas et al. 1997). The alignment was compared to the one available from the HIV Sequence Database for consistency (Kuiken et al. 2001). Regions of indels were kept and treated as sequence ambiguity. Several phylogenetic trees were constructed using the complete genomes and different methods such as neighbour-joining (NJ), maximum parsimony (MP) and maximum-likelihood (ML). The topologies were highly similar to the published lentivirus phylogeny. Individual genes trees were also constructed using the above methods as implemented in PAUP (Swofford 2000). The *env* and *nef* gene trees differed from the "genomic" and other gene trees.

4.2.2 Detection of Adaptive Evolution and Inference of Positive Selection Sites

The HIV-2 and SIVsm datasets were analysed separately to elucidate the differences in the pattern of selective pressure at a molecular level. This analysis is also termed "site-specific" analysis, in that the substitution rate is allowed to vary among sites, but not among lineages (Yang et al. 2000). The datasets were analysed using six codon substitution models (M0, M1, M2, M3, M7 and M8), with the exception of M0, all models allowed rate variation across the genome (as implemented in PAML 3.13). The one-ratio model (M0) was used as a null to test for among site rate variation (Yang 2000). The "neutral" models (M1 and M7) were used as null hypotheses to test for adaptive evolution. The selection, discrete and beta&ω models (M2, M3 and M8) were implemented to account for positive selection. The likelihood ratio tests (LRTs) were used to compare the alternative models to their null. In addition to d_N/d_S ratios (ω), the models also estimated branch lengths of the tree (t), transition/transversion rate ratio (κ) and equilibrium codon frequencies (π). To minimise the effect of recombination, an unrooted phylogeny was used for parameter estimation. Positive selection sites were inferred using Bayesian methods. Sites belonged to $\omega > 1$ rate class, with posterior probability exceeding 90% were considered as candidate positive selection sites. Only candidate sites that were identified by all three "positive selection" models (M2, M3 and M8) were kept. Positive selection sites were numbered using the "HIV/SIV sequence locator tool" (Calef et al. 2001) and mapped onto CTL epitopes (O'Connor et al. 2001).

4.2.3 Detection of Recombination and Inference of Recombinant Sequences

The datasets were analysed using PIST to detect recombination (Worobey 2001). For each dataset (HIV-2 and SIVsm), 1000 replicates were simulated assuming nucleotide substitution model HKY85 and along a ML tree. Among site rate variation was described by a gamma distribution with the shape parameter α set to 0.1. The gamma distribution was approximated into four discrete categories. Codon usage and transition/transversion ratio biases were accounted for using the parameter estimates (π, and κ, after rescaling κ) obtained using ML. The intensity of recombination was determined by ISI and its significance was assessed by the probability score. To minimise the effect of selection, the analysis of third codon positions was conducted. The analysis of all codon positions was a test for robustness. Only datasets with significant probability scores (a strong signal of recombination) were analysed using PLATO for the identification of possible recombinant regions (Worobey. 2001; Grassly and Holmes. 1997). The above assumptions and a minimal sliding window size of five nucleotides were assumed in this study.

4.2.4 Estimation of d_N/d_S ratios (ω) for Different Lineages

This analysis is sometimes referred to as "lineage-specific" analysis, as the models allowed two or more d_N/d_S ratios for specified lineages, though substitution rates were averaged over all the sites (Yang 1998). The HIV-2 and SIVsm genomes were pooled into a large dataset to compare HIV-2 with SIVsm lineages using ML models. Three lineage specific-models were used, two two-ratios (R2a and R2b) and one three-ratio (R3). Two ω ratios were estimated

for HIV-2 and SIVsm clades, with HIV-2 sequences as the foreground lineages (R2a, Figure 4.1). In model R2b, different d_N/d_S ratios were estimated for the internal branch and the rest of the tree (Figure 4.2). The three-ratio model (R3) assumed different rates for the internal branch, the progenitor (SIVsm) and the "daughter" HIV-2 clades. The LRT comparing M0 (one-ratio model) with R2a was to test rate difference between the progenitor and the "daughter" clade. The LRT of M0 with R2b was a test for rate variation in the internal branch when compared to the clades. Since the models were nested, the two-ratios models (R2a and R2b) could be compared with the three-ratio models with one degree of freedom. The transition/transversion rate ratio and the empirical codon frequencies were also estimated for these analyses.

Figure 4.1- Topology used in model R2a The HIV-2 (foreground lineages are coloured red)

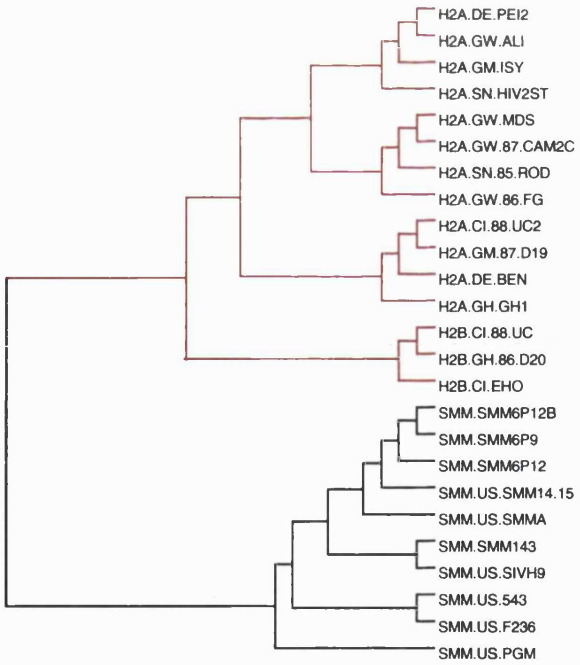
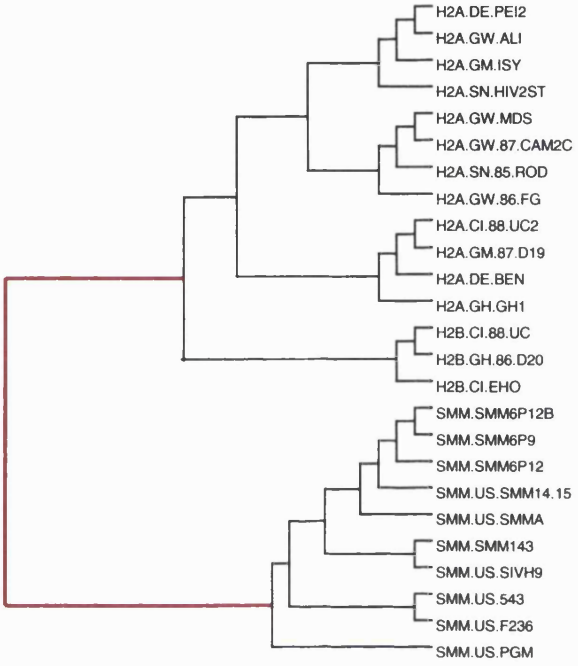


Figure 4.2-Tree topology used in model R2b The internal branch is coloured red



4.2.5 Estimation of ω Assuming Lineage and Site Specific Evolution

The combined dataset was analysed using three codon substitution models that accounted for lineage and site- specific rate variation (Model A, B and D). Model A and B were implemented to detect adaptive evolution in lineages of interest (Yang and Nielsen 2002). Hence, they are ideal models to describe the evolution of HIV-2 after its arrival in human. It is probable that a burst of evolution has occurred in HIV-2 lineages and these lineages are classed as "foreground" branches, with the rest classed as "background" branches. These models assumed four rate classes, ω_0, ω_1, $\omega_0 \rightarrow \omega_2$ and $\omega_1 \rightarrow \omega_2$. The first two rate classes

(ω_0 and ω_1) were assumed to be the same across all lineages and the other two classes permitted some sites in the foreground lineages (with ω_0 or ω_1) to change to ω_2. In Model A, ω_0 and ω_1 were fixed to 0 and 1 respectively. Thus, positive selection could only be detected in foreground branches (HIV-2). Model B allowed ω_0 and ω_1 to be estimated, which could exceed 1. Hence positive selection could be operating at some sites across all lineages (HIV-2 and SIVsm) and other sites could be evolving by adaptive evolution in foreground branches only (Yang and Nielsen 2002). The LRT comparing M1 (neutral) with Model A was a test for positive selection in HIV-2 lineages, after cross-species transmission. The LRT comparing Model B with M3 (discrete) was a test for positive selection in HIV-2, while allowing adaptive evolution to operate in all lineages (Bielawski and Yang 2003). Model D differed from Model A and B in that it assumed four different ω classes, ω_0, ω_1, ω_A and ω_B. The ω_0 and ω_1 classes were estimated from data and assumed to be equal for the entire phylogeny. A fraction of sites were allowed to evolve at two different rates, in SIVsm (ω_A) clades and in HIV-2 (ω_B) lineages. The LRT comparing Model D with M3 was a test for change of selective pressure at certain sites, which are not necessarily positively selected (Bielawski and Yang *submitted*). All models have multiple optima and four different initial values (0.01, 0.1, 3 and 5) were used for parameter estimations.

4.3 RESULTS

4.3.1 Positive Selection Detected in HIV-2 Genes

Parameter estimates under the six codon substitution models were shown in Table 4.1 as single gene and genomic analysis. Variation in selective pressure across the genome was evident, even when d_N/d_S ratios were estimated as an average across all sites and lineages (M0). In highly conserved genes, such as *gag* and *pol*, the averaged ω was very small, 0.13 and 0.08 respectively. In less conserved genes, such as *env* and *nef*, the averaged ω was much higher, 0.31 and 0.36 respectively. Under models that allowed rate variation, similar patterns of selective pressure (measured by d_N/d_S ratios) were detected in all five major genes (*gag*, *pol*, *vif*, *env* and *nef*). Most of the sites in these genes were evolving by strong purifying selection, while the majority of the remaining sites were evolving under relaxed selective constraint (Table 4.1, the notations are the same as before). However, LRTs indicated that each gene possessed a small proportion of sites evolving by positive selection (Table 4.2). Adaptive evolution was also detected in two small genes *vpr* and *vpx*. The analysis of the genome as a "super gene" sequence with all the coding regions concatenated (overlapping regions were excluded) was conducted as a test for the robustness. Results of this analysis were highly similar to that of the single gene analyses

Table 4.1 Parameter estimates for HIV-2 genes.

		Parameter estimates under different models					
Gene	L_C	M0 (one-ratio)	M1 (neutral)	M2 (selection)	M3 (discrete)	M7 (beta)	M8 (beta&ω)
gag	469	ω = 0.13	p_0 = 0.60, ω_0 = 0	p_0 = 0.61, ω_0 = 0	p_0 = 0.72, ω_0 = 0.01	B(0.15, 0.59)	B(0.29, 1.18)
			p_1 = 0.40, ω_1 = 1	p_1 = 0.34, ω_1 = 1	p_1 = 0.21, ω_1 = 0.26		p_0 = 0.95
				p_2 = 0.04, ω_2 = 4.81	**p_2 = 0.07, ω_2 = 1.43**		**p_1 = 0.05, ω = 1.99**
pol	919	ω = 0.08	p_0 = 0.68, ω_0 = 0	p_0 = 0.70, ω_0 = 0	p_0 = 0.70, ω_0 = 0.05	B(0.22, 1.23)	B(0.33, 1.14)
			p_1 = 0.32, ω_1 = 1	p_1 = 0.29, ω_1 = 1	p_1 = 0.26, ω_1 = 0.52		p_0 = 0.99
				p_2 = 0.01, ω_2 = 4.20	**p_2 = 0.04, ω_2 = 2.14**		**p_1 = 0.01, ω = 3.49**
vif	133	ω = 0.22	p_0 = 0.50, ω_0 = 0	p_0 = 0.50, ω_0 = 0	p_0 = 0.61, ω_0 = 0.06	B(0.19, 0.29)	B(0.21, 0.31)
			p_1 = 0.50, ω_1 = 1	p_1 = 0.47, ω_1 = 1	p_1 = 0.36, ω_1 = 0.95		p_0 = 0.97
				p_2 = 0.03, ω_2 = 2.12	**p_2 = 0.03, ω_2 = 1.79**		**p_1 = 0.03, ω = 2.05**
vpr	56	ω = 0.62	p_0 = 0.46, ω_0 = 0	p_0 = 0.60, ω_0 = 0	p_0 = 0.67, ω_0 = 0.03	B(0.16, 0.26)	B(0.29, 0.78)
			p_1 = 0.54, ω_1 = 1	p_1 = 0.31, ω_1 = 1	p_1 = 0.23, ω_1 = 0.55		p_0 = 0.91
				p_2 = 0.09, ω_2 = 3.07	**p_2 = 0.10, ω_2 = 2.10**		**p_1 = 0.09, ω = 2.11**
env	767	ω = 0.31	p_0 = 0.64, ω_0 = 0	p_0 = 0.63, ω_0 = 0	p_0 = 0.74, ω_0 = 0.03	B(0.10, 0.43)	B(0.21, 0.48)
			p_1 = 0.36, ω_1 = 1	p_1 = 0.35, ω_1 = 1	p_1 = 0.24, ω_1 = 0.69		p_0 = 0.98
				p_2 = 0.02, ω_2 = 10.04	**p_2 = 0.02, ω_2 = 4.96**		**p_1 = 0.02, ω = 4.02**
nef	203	ω = 0.36	p_0 = 0.53, ω_0 = 0	p_0 = 0.51, ω_0 = 0	p_0 = 0.75, ω_0 = 0.07	B(0.23, 0.74)	B(0.26, 0.79)
			p_1 = 0.47, ω_1 = 1	p_1 = 0.46, ω_1 = 1	p_1 = 0.22, ω_1 = 0.43		p_0 = 0.97
				p_2 = 0.03, ω_2 = 4.42	**p_2 = 0.03, ω_2 = 2.50**		**p_1 = 0.03, ω = 2..28**
genome	2574	ω = 0.21	p_0 = 0.79, ω_0 = 0	p_0 = 0.79, ω_0 = 0	p_0 = 0.83, ω_0 = 0.03	B(0.15, 0.36)	B(0.16, 0.70)
			p_1 = 0.21, ω_1 = 1	p_1 = 0.20, ω_1 = 1	p_1 = 0.15, ω_1 = 0.68		p_0 = 0.98
				p_2 = 0.01, ω_2 = 21.18	**p_2 = 0.02, ω_2 = 5.66**		**p_1 = 0.02, ω = 5.28**

Table 4.2 Likelihood ratio statistics (2 Δλ) for site-specific analysis

Gene	M0 vs. M3 ($\chi^2_{0.01,\,4} = 13.28$)	M1 vs. M2 ($\chi^2_{0.01,\,2} = 9.21$)	M7 vs. M8 ($\chi^2_{0.01,\,2} = 9.21$)
gag	61.03	22.31	19.56
pol	54.86	32.54	34.91
vif	70.80	30.12	21.91
vpr	44.69	13.43	14.05
env	123.43	56.19	42.47
nef	77.96	63.12	51.15
genome	2499.05	1029.82	991.64

Both the genomic and the single gene analyses have identified the same set of positive selection sites (67 sites genome wide), at a threshold probability of 90%. However, the majority of the sites (42 positive selection sites) were located in *env* and clustered in the primary sequence. It is possible that the variability of these regions was generated by recombination instead of selection.

Table 4.3 Sites identified as evolving by positive selection under M2

Gene	Single gene analysis	Genomic analysis
gag	146V	146V
pol	**278D**, **362Q**, 366R,	278D, 362Q
vif	31V, **36R**, 61D, 63R,	36R, 63R
vpr	**37I**, 55A, 60I, 77R,	**60I**, **77R**

env	62D, **87V**, 178K, **232T**, **308R**, **362K**, 446S, 640S, 817A	87V, 178Y, **362K**, 446S, 640S, **817A**,
nef	179E, **206D**	179E, 206D

The sites were numbered using HIV/SIV Sequence Locator (Kuiken et al. 2001). Positive selection sites with posterior probability $P \geq 95\%$ are in bold. Those located in possible recombinant regions were excluded. 16

4.3.2 Positive Selection Detected in SIVsm

Estimates indicated that the pattern of selective pressure was not uniform throughout the genome, with positive selection only acting on certain genes (Table 4.4). SIVsm genes appeared to be more conserved, with fewer sites evolving by positive selection and/or relaxed functional constraint. My results suggest that in *gag*, *vif* and *nef* strong purifying selection was acting on most of the sites and the remaining sites were evolving by weak functional constraint.

Table 4.4 Parameter estimates for SIVsm genes

		Parameter estimates under different models					
Gene	L_C	M0 (one-ratio)	M1 (neutral)	M2 (selection)	M3 (discrete)	M7 (beta)	M8 (beta&ω)
gag	493	$\omega = 0.12$	$p_0 = 0.70$, $\omega_0 = 0$	$p_0 = 0.60$, $\omega_0 = 0$	$p_0 = 0.72$, $\omega_0 = 0.05$	$B(0.20, 0.59)$	$B(0.29, 0.98)$
			$p_1 = 0.30$, $\omega_1 = 1$	$p_1 = 0.02$, $\omega_1 = 1$	$p_1 = 0.23$, $\omega_1 = 0.59$		$p_0 = 0.95$
				$p_1 = 0.38$, $\omega_1 = 0.41$			$p_1 = 0.05$, $\omega = 0.79$
pol	919	$\omega = 0.06$	$p_0 = 0.68$, $\omega_0 = 0$	$p_0 = 0.68$, $\omega_0 = 0$	$p_0 = 0.81$, $\omega_0 = 0.09$	$B(0.24, 1.02)$	$B(0.33, 0.74)$
			$p_1 = 0.32$, $\omega_1 = 1$	$p_1 = 0.02$, $\omega_1 = 1$	$p_1 = 0.17$, $\omega_1 = 0.52$		$p_0 = 0.99$
				$p_2 = 0.31$, $\omega_2 = 0.32$	**$p_1 = 0.02$, $\omega_1 = 99.90$**		**$p_1 = 0.02$, $\omega = 99.9$ 9**
vif	133	$\omega = 0.23$	$p_0 = 0.50$, $\omega_0 = 0$	$p_0 = 0.50$, $\omega_0 = 0$	$p_0 = 0.61$, $\omega_0 = 0.06$	$B(0.19, 0.29)$	$B(0.21, 0.31)$
			$p_1 = 0.50$, $\omega_1 = 1$	$p_1 = 0.47$, $\omega_1 = 1$	$p_1 = 0.36$, $\omega_1 = 0.95$		$p_0 = 0.97$
env	771	$\omega = 0.46$	$p_0 = 0.77$, $\omega_0 = 0$	$p_0 = 0.73$, $\omega_0 = 0$	$p_0 = 0.74$, $\omega_0 = 0.08$	$B(0.02, 0.05)$	$B(0.02, 0.05)$
			$p_1 = 0.23$, $\omega_1 = 1$	$p_1 = 0.24$, $\omega_1 = 1$	**$p_1 = 0.23$, $\omega_1 = 1.09$**		$p_0 = 0.96$
				$p_2 = 0.03$, $\omega_2 = 5.35$	**$p_2 = 0.03$, $\omega_2 = 5.60$**		**$p_1 = 0.04$, $\omega = 5.04$**
nef	203	$\omega = 0.39$	$p_0 = 0.71$, $\omega_0 = 0$	$p_0 = 0.74$, $\omega_0 = 0$	$p_0 = 0.75$, $\omega_0 = 0.00$	$B(0.03, 0.07)$	$B(0.01, 1.27)$
			$p_1 = 0.29$, $\omega_1 = 1$	$p_1 = 0.00$, $\omega_1 = 1$	$p_1 = 0.25$, $\omega_1 = 1.35$		$p_0 = 0.76$
				$p_2 = 0.26$, $\omega_2 = 1.34$			$p_1 = 0.04$, $\omega = 1.34$
genome	2574	$\omega = 0.24$	$p_0 = 0.76$, $\omega_0 = 0$	$p_0 = 0.76$, $\omega_0 = 0$	$p_0 = 0.74$, $\omega_0 = 0.00$	$B(0.01, 0.14)$	$B(0.06, 0.20)$
			$p_1 = 0.24$, $\omega_1 = 1$	$p_1 = 0.22$, $\omega_1 = 1$	$p_1 = 0.25$, $\omega_1 = 0.78$		$p_0 = 0.98$
				$p_2 = 0.01$, $\omega_2 = 6.85$	**$p_2 = 0.01$, $\omega_2 = 5.78$**		**$p_1 = 0.01$, $\omega = 5.32$**

Table 4.5 Likelihood ratio statistics (2 $\Delta\lambda$) for hypothesis testing

Gene	M0 vs. M3 ($\chi^2_{0.01,\,4} = 13.28$)	M1 vs. M2 ($\chi^2_{0.01,\,2} = 9.21$)	M7 vs. M8 ($\chi^2_{0.01,\,2} = 9.21$)
gag	17.09	0.13	1.36
pol	15.21	0.54	1.29
vif	21.07	0.79	1.91
env	225.8	56.24	56.74
nef	34.70	1.56	1.56
genome	2499.05	1029.82	991.64

Table 4.6 Sites identified as evolving by positive selection under M2

Gene	Single gene analysis	Genomic analysis
env	279T, **287S**, 288N, 291I, 301T, **312S**, 321S, 341T, **344V**, 347S, 406T	279T, 287S, 288N, **291I**, 301T, 312S, **321S**, 341T, 344V, 347S, 406T

Estimation under M3 suggested that all the sites in these genes could be approximated into two discrete rate classes, conserved and "variable" (with $\omega \geq 1$). The neutral" models (M1 and M7) was not rejected, as positive selection did not seem to operate in these genes (Table 4.5). A slightly different pattern of selective pressure was observed in *pol*, ω estimates under M3 and M8 suggested that almost all of the sites (~99.8%) were highly conserved with a d_N/d_S of 0.09 (Table 4.4). Although, positive selection was detected in a very small proportion of sites (~0.02%), yet no positive selection sites were inferred by the Bayesian approach. The high ω estimates for M3 and M8 could be an artefact created by their location near boundary of the class. Hence, no significant support for positive selection was obtained. Interestingly, these results indicated that only *env* was evolving by positive selection in SIVsm (Table 4.4). At 90% cut off, only 11 sites were identified as positive selection sites in SIVsm genome (Table 4.6). Parameter estimates of the full-length genomes were highly similar to that of the single gene analyses.

4.3.2 Significant Support for Recombination in HIV-2 but not in SIVsm

The probability scores indicated at least some of the sequence variability observed within the HIV-2 was due to recombination. At a probability score less than 0.001, HIV-2 genomes were estimated to have an ISI of 0.51. The ISI was an index used to measure the intensity of recombination that could range from zero (no recombination) to one (many recombinations). An ISI of 0.51 suggested a moderate influence of recombination in the sequence evolution, which was statistically significant. The SIVsm genomes had a slightly lower ISI of 0.45 at a probability score less than 0.29. These estimations indicated possible recombination events in the genome of SIVsm, however the result was not statistically significant. Similar

estimates were obtained for the analyses of third codon positions and all codons. PLATO output all the possible recombinant regions and the significance of which were determined by Z-values. Z-value assessed the departure of likelihood from the expected for a particular window size. The Bonferroni-corrected significance was calculated by PLATO to account for multiple window sizes. For these datasets a Bonferroni-corrected significance for α of 0.05 showed that any Z-values greater than 4.21 were considered to be significant. Six possible recombinant regions were identified to be statistically significant in HIV-2. However, one region (4668-4863) was not sensible as this region was concatenated by joining the last non-overlapping reading frame of *vpx* to the beginning of *env*. Four regions were mapped to *env*, where 42 positive selection sites clustered in the primary sequence. Hence these sites were not considered to be under positive selection. The last region was mapped to the beginning of *nef*, where three positive selection sites were excluded. Although no significant support for recombination was obtained for SIVsm using PIST, PLATO identified four possible recombinant regions in SIVsm *env*. One region (6540-6548) was supported by a Z-value identical to the threshold value and was not considered as a recombinant region. The remaining three regions had highly significant statistical support for recombination. As these regions were identified on the basis that they have a different phylogenetic relationship to the rest of the sequence, I reconstructed the topologies of these regions and compared them to the gene tree. Interestingly the only difference between the recombinant regions and the gene tree was the positions of the PBJ clones. As the genetic distance between the clones were very small (measured by branch lengths 0-0.005), the order of the topology for this part of the tree might be unimportant. Hence, no significant support was obtained for recombination in SIVsm.

4.3.3 Different d_N/d_S Estimates for Different Part of the Phylogeny

My results suggest that different lineages were evolving at different rates. The one-ratio model was rejected with statistical significance when compared to models permitting two different rates for different lineages (Table 4.7). Model R2a allows selective pressure to differ between the progenitor and HIV-2 fitted the data much better than M0. The internal lineage (immediately post zoonosis) was permitted to evolve at a different rate from the rest in model R2b. Parameter estimates under model R2b suggested a slow rate for the internal branch ($\omega = 0.05$) and a relatively higher ω ratio for the clades ($\omega = 0.21$). This rate difference was found to be statistically significant, with a P-value < 0.0001 (Table 4.7). Estimates of d_N/d_S under model R3 indicated a slightly higher rate of nonsynonymous substitution among the SIVsm lineages ($\omega = 0.25$) as compared to HIV-2 clades ($\omega = 0.21$). The ω estimated for SIVsm clades under model R3 was much higher than in R2a (0.25 and 0.15 respectively). This appears to be an artefact of estimating ω_{SIVsm} as an average of the internal branch and the SIVsm clades, resulted in an overall low d_N/d_S ratio for SIVsm ($\omega = 0.15$). In model R3, the estimate of ω_{SIVsm} is improved by the addition of $\omega_{internal}$ as a new parameter (Table 4.7). Hence, model R2a was rejected with significant support when compared with R3. To test if selective pressure in SIVsm and HIV-2 differed from that immediately post zoonosis, I compared model R2b with R3. LRT showed that model R3 fitted the data significantly better than R2b (P-value = 0.01). Overall, the likelihood of model R3 was significantly better than other models indicating substantial among lineage rate variation.

Table 4.7 Parameter estimates and likelihood ratio test statistics (2 $\Delta\lambda$)

Models	Tree Length (t)	Parameter Estimates	2 $\Delta\lambda$	P-value
One ratio (M0)	4.77	$\omega = 0.20$	-	-
Two-ratios (R2a)	4.79	$\omega_{HIV-2} = 0.21$, $\omega_{SIVsm} = 0.15$	-	-
M0 vs. R2a ($\chi^2_{0.01, 1} = 6.63$)	-	-	22.30	<0.0001
Two-ratios (R2b): internal /clades	4.95	$\omega_{internal} = 0.05$, $\omega_{clade} = 0.21$	-	-
M0 vs. R2b ($\chi^2_{0.01, 1} = 6.63$)	-	-	183.88	<0.0001
Three-ratios (R3): internal / HIV-2 /SIVsm	4.96	$\omega_{internal} = 0.05$, $\omega_{HIV-2} = 0.21$ $\omega_{SIVsm} = 0.25$	-	-
R3 vs. R2a ($\chi^2_{0.01, 1} = 6.63$)	-	-	167.80	<0.0001
R3 vs. R2b ($\chi^2_{0.01, 1} = 6.63$)	-	-	6.22	0.01

Note. – An unrooted phylogeny was used for parameter estimation under M0, Model R2b and R3. A rooted tree was used for estimates under model R2a (see Figure 4.1 and 4.2).

4.3.4 Positive Selection Detected in Foreground and Background Lineages

Estimates under Model A showed that across the entire phylogeny, 54% of the sites were highly conserved ($\omega_0 = 0$) and 43% of sites were evolving "neutrally" ($\omega_1 = 1$).
The results also indicated that a small proportion of sites in HIV-2 lineages (~3%) were evolving by strong positive selection, with ω as high as 14.77 (Table 4.8). Likelihood ratio statistics showed that model A fitted the data significantly better than M1, by permitting 1.4% of the sites from ω_0 and 1.1% of the sites from ω_1 to change to ω_2 rate class (Table 4.9). At a critical threshold of 50%, 62 positive selection sites were identified in HIV-2 genome. These sites were also identified in the sites-specific analysis, but with much higher posterior probabilities (i.e. $P \geq 90\%$). Parameter estimates under model B indicated that across all lineages, the majority of the sites were either highly conserved ($\omega = 0.03$) or evolving by relaxed functional constraint ($\omega = 0.70$). Positive selection was only detected in foreground branches, with 3% of the sites in HIV-2 lineages identified as positive selection sites (Table 4.8). At a critical threshold of 50%, model B identified fewer sites than model A. Model D allows a proportion of sites to evolve at different rates ω_A (for SIVsm lineages) and ω_B (for HIV-2). Estimation under model D suggested that positive selection is operating in both SIVsm and HIV-2 lineages (Table 4.8). Likelihood ratio test supported this hypothesis with statistical significance (Table 4.9). The sites estimated as positive selection sites in model D are also identified in site-specific analyses of SIVsm and HIV-2

Table 4.8 Parameter estimates and likelihood score under Branch-site models

Model	p	Parameter Estimates	ℓ
Site specific			
Neutral (M1)	1	$p_0 = 0.54, p_1 = 0.46$	-52116.77
Discrete (M3), $k = 2$	3	$p_0 = 0.80, \omega_0 = 0.15$	-50381.42
		$p_0 = 0.20, \omega_2 = 1.61$	
Discrete (M3), $k = 3$	5	$p_0 = 0.74, \omega_0 = 0.03$	-50379.42
		$p_0 = 0.23, \omega_1 = 0.54$	
		$p_0 = 0.03, \omega_2 = 3.11$	
Branch-site			
Model A	3	$p_0 = 0.54, \omega_0 = 0$	-51594.83
		$p_1 = 0.43, \omega_1 = 1$	
		$p_{(2+3)} = 0.03, \omega_2 = 14.77$	
Model B	5	$p_0 = 0.74, \omega_0 = 0.03$	-50403.26
		$p_1 = 0.23, \omega_1 = 0.70$	
		$p_{(2+3)} = 0.03, \omega_2 = 6.33$	
Model D	6	$p_0 = 0.72, \omega_0 = 0.03$	-50304.26
		$p_1 = 0.24, \omega_1 = 0.57$	
		$p_2 = 0.04, \omega_A = 2.42, \omega_B = 4.31$	

Table 4.9 Likelihood ratio test statistics

	$2\Delta\lambda$	d.f.	P-value
LRT of variable ω between SIVsm and HIV-2 clades			
M1 vs. Model A	1043.88	2	< 0.0001
M3 ($k = 2$) vs. Model B	47.68	2	< 0.0001
M3 ($k = 3$) vs. Model D	150.32	1	< 0.0001

4.4 Discussion

4.4.1 Evidence of Adaptive Evolution Operating in HIV-2 and SIVsm

Many researches mainly focused on the evolution of HIV-1 and relatively less studies were conducted on HIV-2. As HIV-1 infections are more prevalent and lethal in comparison to HIV-2, it is likely to receive more research interest. However, the introduction of SIVsm into human is an excellent model to study evolution post zoonosis. After the cross-species jump, both progenitors are unlikely to experience the same intensity of immune selective pressure. Hence it is interesting to elucidate the viral adaptive response to this change. Previous single gene analyses on *gp120*, and *nef* of HIV-2 have indicated a possible role for adaptive evolution (Shpaer and Mullins 1993). However, the methods employed by the authors, (i.e. NS/S >1) was not suitable to detect positive selection. I expanded upon the previous study by analysing the entire genome of HIV-2 using a maximum likelihood approach. This study of the two most prevalent subtypes (A and B) was to identify sites under continual selection since its arrival in human. Site-specific analyses have indicated that most HIV-2 genes (*gag*, *pol*, *vif*, *env,* and *nef*) were evolving by positive selection. My analyses of the two small genes *vpr* and *vpx* have also indicated a role for adaptive evolution. To summarise, my study is the first to suggest that like HIV-1, positive selection promotes sequence divergence in HIV-2 genes.

Previous studies on the rates of amino acid changes in *env* of SIVsm indicated no role for positive selection (Shpaer and Mullins 1993). However, my results have suggested that a small portion of the sites in SIVsm *env* is evolving by adaptive evolution. These sites were mapped to CTL epitopes in naturally and experimentally infected sooty mangabeys (Kaur et al. 2000). My findings were consistent with the notion that SIV proteins are immunogenic

within the natural host (Kaur et al. 2001). The immunogenicity of *env* is thought to trigger a broadly directed CTL specific response in mangabeys, which is sustained during the course of infection (Courgnaud et al. 1998; Kaur et al. 2001). This continual selective pressure is likely to result in an accumulation of replacement substitutions as compared to synonymous changes (i.e. $\omega > 1$). Hence regions with an excess of nonsynonymous substitutions could be considered as antigenic regions. I have identified several such regions in SIVsm *env*, indicating the importance of CTL recognition in generating sequence variation. Viral epitope specific CTL responses in non-pathogenic infections of SIVsm may not be as rare as previously anticipated. Longitudinal experiment study has suggested that Nef specific CTL responses tend to persist throughout the course of avirulent infections of SIVsm (Kaur et al. 2001). My results supported this notion, as sites with ω close to 1 in the *nef* gene of SIVsm were mapped to known CTL epitopes and may represent possible CTL escape mutations.

4.4.2 The Impact of Recombination

As mentioned in Chapter 1, recombination occurs frequently in primate lentiviruses and generates a high level of sequence diversity. However, this sequence divergence could be misinterpreted as signals of positive selection (Anisimova et al. 2003). Recombinant data tends to have a complex evolutionary history best described by a set of correlated trees. Hence, assuming one phylogeny may lead to false detection of sites under selection, as recombinant sequences could be interpreted as signals of among site rate variation. Therefore, it is important to elucidate the impact of recombination on my data. Falsely detected positive selection sites tend to cluster in the primary sequence, as observed in the four regions of HIV-2 genome (see results). Positive selection sites located in these regions

were excluded as false positives. It is worth noting that these possible hybrid regions represent intra-subtype recombination, as the data was free of inter-subtype recombination. Intra-subtype recombination is often ignored in evolutionary analyses of viral data (Nielsen and Yang 1998). As recurrent selection could generate patterns of variation similar to that of recombination, one must minimise such possibilities. In this study, I analysed both third codon positions and all codon positions to ensure the robustness of my results. Also, in the inference of hybrid regions, I assumed among site substitution rate heterogeneity. After minimising the effect of selection, recombination was still detected in the *env* and *nef* genes of HIV-2. My findings support the notion that recombination may occur frequently in the Env protein of HIV-2 (McVean et al. 2002). Interestingly, no significant support for recombination was obtained for SIVsm genomes. As discussed in Chapter 1, SIVsm recombines with different lineages of SIVs to form new hybrid viruses. Hence, the genome of SIVsm is very tolerant of such genetic shuffling. Therefore, it is probable that recombination should occur within the SIVsm lineage to generate genetic diversity. To evaluate the impact of such recombination, one needs an extensive dataset of SIVsm genomes sampled from different populations. The lack of such data greatly hinders our understandings of recombination and the genetic flexibility of primate lentiviruses.

4.4.3 Lineage Specific Variation in Selective Pressure

Since the arrival of HIV-2 into human population, it has undergone many evolutionary changes, which could be interpreted as the genetic divergence observed between HIV-2 and its progenitor (Hahn et al. 2000). This burst of evolution is thought to be a direct response to the elevated selective pressure experienced by the virus post zoonosis. Hence, one would

expect a significant increase in the rate of fixation of replacement substitutions immediately postdate the zoonotic event. However, my results suggested a low ω just after the transmission. It is possible that during the initial introduction, a few genetic changes were sufficient for the virus to escape immune detection. These changes may not have to be amino acid altering. For example, HIV-1 was able to escape CTL recognition with synonymous substitutions (da Silva and Hughes 1998). Also, it is likely that the virus can adapt to a foreign host with minimal genetic alterations, because zoonotic transmissions occur readily in freely interacting primates (see Chapter 1). Hence, a low ω might be observed immediately postdating the transmission event. My results indicated that after the emergence of HIV-2, a higher rate of nonsynonymous substitutions was observed for HIV-2 lineages. This accelerated rate could reflect an evolutionary race between host and pathogen (see Chapter 2). It is interesting that SIVsm lineages have a slightly higher ω than HIV-2 clades. However, as SIVsm can replicate to a high level in its natural host and many viral proteins might still retain their immunogenicity throughout infection (Holmes 2001; Kaur et al 2000). It is probable that this high ω reflects the sustained selective pressure exerted by SIVsm specific CTL response, which is observed during the infection (Kaur et al. 2001).

4.4.4 Evidence of Adaptive Evolution in SIVmac and SIVsm Lineages

Site-specific analysis of the HIV-2/SIVsm dataset indicated that when averaged across the entire phylogeny, a fraction of sites were evolving by adaptive evolution (see M3, Table 4.9). However, it is more interesting to identify sites, which became positively selected since the arrival of HIV-2 in human. These amino acids represented adaptation at a molecular level to

the changes in host environment. Branch-site models (A and B) have detected positive selection acting on a proportion of the sites (62 and 56 sites, respectively) in HIV-2 lineages. The majority of these sites were located in possible hybrid regions of HIV-2 and hence were excluded from the results. The remaining sites were mapped to HIV-2 specific CTL epitopes indicating possible adaptive evolution promoted by immune recognition. However, model B did not detect any site evolving by adaptive evolution across the entire phylogeny (ω_1 = 0.70). Estimates under model D suggested that a fraction of sites (4%) were evolving by positive selection in both SIVsm and HIV-2 lineages ($\omega_A = 2.42$, $\omega_B = 4.31$). These results were potentially contrasting. However, Bayesian approaches have identified the same subset of sites in model D as positive selection sites, indicating a strong positive selection signal at these sites. Model B is more conservative than model D, in that it requires $p_{(2+3)}$ to be drawn from $\omega_0 \rightarrow \omega_2$ and $\omega_1 \rightarrow \omega_2$ rate classes. Also to identify positive selection, the estimates ω_1 and ω_2 have to exceed one. Hence, it is possible that if a fraction of sites evolving by positive selection in ω_1 were drawn to account for adaptive evolution in foreground lineages (HIV-2), the remaining positive selection sites were undetected, due to a reduction of power. To summarise, my results in sites-specific and branch-site specific analyses indicated that adaptive evolution was operating in both HIV-2 and SIVsm genes. However, more genes were evolving by positive selection in HIV-2 than SIVsm, probably reflecting adaptation to new host.

Chapter 5 THE EVOLUTION OF A SIMIAN IMMUNODEFICIENCY VIRUS (SIVMAC) POST ZOONOSIS: THE USE OF SIVMAC INFECTION OF MACAQUES AS A NONHUMAN MODEL.

5.1 The Emergence of a New Immunodeficiency Virus in Macaques

Prior to the outbreak of HIV-2 infection, the development of immunodeficiency related disorders together with an unusual clustering of lymphomas were observed in a group of rhesus macaques at the New England Regional Primate Research Centre (Hunt et al. 1983; Letvin et al. 1983). Subsequent investigation resulted in the isolation of a T-cell tropic retrovirus virus that caused AIDS like symptoms. The virus was designated SIVmac. Pathogenesis studies revealed that SIVmac induced fatality in challenged captive Asian macaques (Doolittle 1989; Hirsch et al. 1989). As discussed in Chapter 1, SIVmac was suspected to be the progenitor of HIV-2 until the isolation of SIVsm in sooty mangabeys. It appears to be that SIVmac is another example of accidental zoonotic transmission of SIVsm.

All primate lentiviruses are capable of infecting and killing T-helper ($CD4^+$) cells (i.e. they are all T-tropic retroviruses) (Shpaer and Mullins 1993). However, it appeared to be that in its natural hosts, the virus is avirulent (reviewed in Chapter 1). However, after zoonosis the emerged virus became more pathogenic than its progenitor. In general, SIV and HIV only induces immunodeficiency and AIDS like symptoms after a long incubation period, where the virus remains latent. However, a particularly virulent strain of SIVmac was isolated (SIVpbj14) after nine passages of the parental SIVsm through the macaque. This strain was able to replicate in peripheral blood mononuclear cells (PBMC), whereas efficient replication of most SIV and HIV was dependent on T-helpers. This key difference resulted in severe illness, such as gastroenteritis, which was not directed related to immunodeficiency and the animal was killed within ten days (Kirchhoff et al. 1999). Positive selection was unlikely to be operating in these lethal clones, as their swift elimination of the host severely hindered their spread. Although the virulence of the strains varies dramatically, the basic

clinical symptoms are very similar, characterised by the loss of T-helper cells and the onset of AIDS. Physical changes in the lymphoid tissues and the development organ-specific disease such as encephalitis are also observed in human AIDS (Joag. 2000). Similar members of the chemokine family were used as the co-receptors for envelope protein and host cell interactions. These features make the SIVmac/macaque model a good nonhuman model for studies on HIV-1 virulence and virion-cell interactions.

Despite the similarities in the clinical symptoms and replication mechanism, the genomic organisation and the biology of the viruses (HIV-1 and SIVmac) were quite different. Also many attempts to infect macaque with HIV-1 had failed induce disease, indicating possible cellular and immunological differences between the hosts. The difference observed in the neutralising domains of HIV-1 and SIVmac (with the former appeared to be linear and the latter conformational), resulted in limitations of studies on passive immunity and vaccine design. Differences between human and macaque MHC class I molecules have also been observed. Notably, the requirement of three anchoring residues at position 2, 3 and C-terminus of the peptide was a unique feature of the macaque MHC-1 binding motif. However, both human and macaque MHC-1 molecules appeared to bind *gag* and *env* at similar regions and HLA-A and B-locus homologues were found in macaque and human (Dzuris et al. 2000). In addition, the immune responses in infected macaques closely resemble those observed in human. T-cell mediated elimination of infected cells and suppression of viral replication were observed in both cases. Neutralising antibodies were produced without any prevention in viral replication, as seen in both hosts. Hence there are obvious advantages in using the SIVmac/macaque model for the study of HIV-1 infection. Nonetheless, the difference mentioned above could be one of the few key disadvantages in

using such a nonhuman model. Therefore it is important to evaluate the reliability of using such a nonhuman model to describe evolutionary events after the zoonosis.

The progenitor, SIVsm does not result in any immune deficiency in its natural host, indicating possible sequence changes post zoonosis. As the date of the zoonosis and the natural hosts were reasonably established, it is possible to test for any change in selective pressure post transmission. The estimation of nonsynonymous (replacement)/synonymous substitution rate ratio (d_N/d_S) could be used to detect changes in selective pressure post zoonosis. Sites identified to be under adaptive evolution post zoonosis could contribute to the observed changes in pathogenesis. The same progenitor gave rise to HIV-2 infection, which was less lethal in comparison to SIVmac (with the progression to AIDS occur in one or two years in rhesus macaque and approximately 90 days for pig-tail macaque), suggesting the change in pathogenesis might be associated with the host species. Such difference in virulence could be due to different intensity of selective pressure exerted on the virus by the differences in the immune systems. Hence sites under adaptive evolution in HIV could differ from those identified in SIV. The identification of such sites would contribute to the evaluation of the reliability of SIVmac/macaque model and the immunotherapies approach developed using such models.

5.2 Material and Methods

5.2.1 Data Preparation and Phylogenetic Inference

The complete genomes of 10 SIVsm and 8 SIVmac were obtained from the HIV Sequence Database. The 10 SIVsm sequences comprised of the six clones and four isolates (L09211-L09213, L03295, M83293, M80193, M80194, U72748, X14307, AF077017). All the SIVmac sequences were isolated from macaques infected with SIVsm from sooty mangabeys (AY033233, AY033146, M76764, M19499, Y00277, M16403, U79412, M83293). All overlapping regions were excluded from the analysis including sections of *gag*, *pol*, *vif*, *env*, and all of *tat*, *rev*, *vpx* and LTRs. The alignment was obtained using Clustal X and manually adjusted using GenDoc. The consistency of the alignment was checked against the one available from the HIV Sequence Database (Kuiken et al. 2001). Regions of indels and genetic code ambiguity (e.g. B, D, H, V and N) were treated as sequence ambiguity. The alignments produced were: a) SIVmac genomes, b) SIVmac/SIVsm genomes, and c) SIVmac genes, i.e. *gag*, *pol*, *vif*, *env*, and *nef*. Various methods were used to infer phylogenetic relationships, as implemented in PAUP (Swofford. 2000). All methods (maximum likelihood, maximum parsimony and neighbour joining) produced similar phylogenies with relatively high bootstrap support. These phylogenies agreed with the published SIV topology (Beer et al. 1999).

5.2.2 Detecting Amino Acid Sites Under Positive Selection

Six codon substitution models M0, M1, M2, M3, M7 and M8 were used for parameter estimations. Substitution rate ratios (ω) were estimated as free parameters under all five

models, with the exception of M1 that requires ω to be either 0 or 1(Yang. 2000). The one-ratio model (M0) only allowed one ω class for all the sites, whereas the other models allowed at least two rate classes. The substitution rate ratio could exceed one only in models allowing for positive selection (i.e. M2, M3 and M8). Non-uniform evolution among sites could be tested by comparing M0 and M3 in a likelihood ratio test (LRT). Three, ten and eleven discrete ω rate classes were assumed for parameter estimation under M3, M7 and M8 respectively. LRTs comparing the "positive selection" models with the neutral models (i.e. M1 with M2 and M7 with M8), were tests for adaptive evolution. Other parameters estimated were equilibrium codon frequency (π), tree branch length (t), and transition/transversion rate ratio (κ). An unrooted phylogeny was used for this analysis to minimise the influence of recombination. Posterior probability estimated by Bayesian approaches was used to infer candidate positive selection sites. Candidate positive selection sites were sites from the $\omega > 1$ rate class with posterior probability exceeding 90%. I only present sites identified by all three "positive selection models" (M2, M3 and M8) as candidate positive selection sites. These sites were numbered using the "HIV/SIV sequence locator tool" (Calef et al. 2001) and mapped onto defined CTL epitopes with known restricting MHC-1 loci molecules (O'Connor et al. 2001).

5.2.3 Detecting Possible Recombinant Regions

To elucidate the effect of recombination within the same lineage of SIV (i.e. SIVmac), I used PIST to detect recombination and PLATO to infer possible hybrid regions (Worobey. 2001; Grassly and Holmes. 1997). To minimise the effect of positive selection, I analysed third codon positions only. As a test for robustness, I have also analysed all codon positions. In

this study, 1000 replicates were simulated along a ML tree. A nucleotide substitution model (HKY85) was used to generate sequence diversity. Among site substitution rate heterogeneity was described using a gamma distribution. The distribution has a shape parameter α and was approximated into four discrete categories. Codon usage and transition/transversion ratio biases were also accounted for in this study (see Chapter 4). The ISI determines the level of recombination observed and the significance of which is assessed by the probability score. Possible hybrid regions were inferred using the same assumptions and at a minimum sliding window size of five nucleotides. The likelihood of a recombinant region was assessed by its Z-value exceeding the threshold value.

5.2.4 Detecting Lineage Specific Changes in Selective Pressure

In this part of the analysis, I was interested in the change of selective pressure post the introduction of SIVsm into macaques. Hence the SIVsm and SIVmac genomes were pooled into one large dataset. The null hypothesis assumes one ω ratio for all lineages. The null is tested against models permitting two different d_N/d_S ratios for different parts of the phylogeny (Yang 1998). Model R2a assumed two independent ω ratios for SIVsm and SIVmac lineages. The tree was rooted at the internal branch i.e. the branch connecting the two clades (resembles Figure 4.1). The LRT comparing model R2a with M0 is a test for variation in selective constraint between the two clades. Model R2b also allowed two independent d_N/d_S ratios, where the internal branch was permitted to evolve at a different rate from the rest of the clades (similar to Figure 4.2). The comparison of model R2b with M0 is a test for variation in selective pressure between the clades and the lineage immediately after the cross-species jump (the internal branch). The two-ratio model R2a could be extended into a more

general model assuming three separate ω ratios: one for the branch immediately postdating zoonosis, a second for the SIVsm clade and the third for all SIVmac lineages. This three-ratio model (R3) could be compared to model R2a and R2b in a LRT with one degree of freedom. Since model R2b and R3 does not require a rooted tree, estimations under these models were obtained using an unrooted phylogeny.

5.2.5 Detecting Lineage Specific Changes in Selective Constraint at Specific Amino Acids

To determine if selective constraint has changed at certain amino acids since the zoonosis; I analysed data using models (Model A, B and D) allowing non-uniform evolution among specific lineages at certain amino acids (Yang and Nielsen. 2002; Bielawski and Yang 2003). In the branch-site models (Model A and B), substitution rate is assumed to differ among sites, but at a subset of sites selective pressure is allowed to change in foreground branches (or SIVmac lineages). The sites permitted to change in the SIVmac lineages could have a ω exceeding one and hence can account for positive selection. This subset of sites could be drawn from either ω_0 or ω_1 rate classes and thus creating two new rate classes, $\omega_0 \rightarrow \omega_2$, and $\omega_1 \rightarrow \omega_2$. The two original rate classes, ω_0, and ω_1 were assumed to be the same across the entire evolutionary history. As mentioned in Chapter 4, model A is an extension of the "neutral" model (M1), with ω_0 and ω_1 fixed at 0 and 1 respectively. As only ω_2 was estimated, this model only allows adaptive evolution in the foreground (SIVmac) lineages. Model B is an extension of the discrete model (M3), with ω_0, and ω_1 estimated as free parameters. Hence, a portion of the sites could be evolving by positive selection across the entire phylogeny and the other sites are under positive selection only in SIVmac lineages.

LRTs were used to compare the alternative models (model A and B) to their nulls (M1 and M3). As mentioned in Chapter4, model D detects divergent selective changes at a proportion of sites, though not necessarily sites under adaptive evolution. Thus the LRT comparing Model D with M3 was a test for divergent selective pressure after zoonosis. Multiple optima were observed in these models and four different initial ω values (0.01, 0.1, 3 and 5) were used for this analysis. Only the parameter estimates obtained under the best likelihoods were presented.

5.3 RESULTS

5.3.1 Adaptive Evolution Detected in SIVmac Genes

Single gene analyses indicated that SIVmac genes were subjected to similar patterns of selective pressure (Table 5.1). LRTs suggested that most of the sites were highly conserved with a small proportion of sites evolving by positive selection (Table 5.2). Estimates under M3 indicated that ω ratios from all sites could be approximated into two discrete classes. For example, in *gag* and *nef*, the estimate for ω_0 is identical to that of ω_1 (i.e. $\omega_0 = \omega_1$). In *pol*, *vif*, and *env* genes, the estimates for ω_1 and ω_2 greatly exceeded one; indicating positive selection is operating at certain sites in these genes (see Table 5.1). Since the estimates for ω_1 and ω_2 classes are both indicative of adaptive evolution, they could be considered as one rate class. Parameter estimates under M2 and M8 also suggested that a fraction of the sites in SIVmac genome is evolving by positive selection. Analysis of the full genomes produced

similar parameter estimates to the single gene analyses and both identified the same set of positive selection sites (46 sites genome wide), at a threshold probability of 90% (Table 5.3).

Table 5.1 Parameter estimates under codon substitution models of variable selective pressure among sites

		Parameter estimates under different models					
Gene	L_C	M0 (one-ratio)	M1 (neutral)	M2 (selection)	M3 (discrete)	M7 (beta)	M8 (beta&ω)
gag	439	$\omega = 0.33$	$p_0 = 0.92$, $\omega_0 = 0$	$p_0 = 0.92$, $\omega_0 = 0$	$p_0 = 0.99$, $\omega_0 = 0.24$	$B(0.01, 0.02)$	$B(0.06, 0.08)$
			$p_1 = 0.08$, $\omega_1 = 1$	$p_1 = 0.07$, $\omega_1 = 1$	**$p_1 = 0.01$, $\omega_2 = 13.32$**		$p_0 = 0.99$
				$p_2 = 0.01$, $\omega_2 = 17.43$			**$p_1 = 0.01$, $\omega = 13.34$**
pol	919	$\omega = 0.15$	$p_0 = 0.95$, $\omega_0 = 0$	$p_0 = 0.94$, $\omega_0 = 0$	$p_0 = 0.96$, $\omega_0 = 0.05$	$B(0.004, 0.02)$	$B(0.02, 0.10)$
			$p_1 = 0.05$, $\omega_1 = 1$	$p_1 = 0.05$, $\omega_1 = 1$	**$p_1 = 0.03$, $\omega_1 = 1.68$**		$p_0 = 0.99$
				$p_2 = 0.003$, $\omega_2 = 20.52$	**$p_2 = 0.003$, $\omega_2 = 21.09$**		**$p_1 = 0.003$, $\omega = 18.28$**
vif	133	$\omega = 0.76$	$p_0 = 0.86$, $\omega_0 = 0$	$p_0 = 0.86$, $\omega_0 = 0$	$p_0 = 0.87$, $\omega_0 = 0.00$	$B(0.03, 0.05)$	$B(0.03, 0.07)$
			$p_1 = 0.14$, $\omega_1 = 1$	$p_1 = 0.07$, $\omega_1 = 1$	**$p_1 = 0.11$, $\omega_1 = 2.19$**		$p_0 = 0.93$
				$p_2 = 0.07$, $\omega_2 = 5.46$	**$p_2 = 0.02$, $\omega_2 = 9.64$**		**$p_1 = 0.07$, $\omega = 5.51$**
env	771	$\omega = 0.70$	$p_0 = 0.86$, $\omega_0 = 0$	$p_0 = 0.86$, $\omega_0 = 0$	$p_0 = 0.93$, $\omega_0 = 0.20$	$B(0.03, 0.08)$	$B(0.05, 0.07)$
			$p_1 = 0.14$, $\omega_1 = 1$	$p_1 = 0.11$, $\omega_1 = 1$	**$p_1 = 0.06$, $\omega_1 = 4.26$**		$p_0 = 0.97$
				$p_2 = 0.03$, $\omega_2 = 12.10$	**$p_2 = 0.01$, $\omega_2 = 32.60$**		**$p_1 = 0.03$, $\omega = 12.75$**
nef	203	$\omega = 1.25$	$p_0 = 0.05$, $\omega_0 = 0$	$p_0 = 0.77$, $\omega_0 = 0$	$p_0 = 0.77$, $\omega_0 = 0.00$	$B(0.33, 0.003)$	$B(0.001, 1.62)$
			$p_1 = 0.95$, $\omega_1 = 1$	$p_1 = 0.00$, $\omega_1 = 1$	**$p_1 = 0.23$, $\omega_1 = 3.39$**		$p_0 = 0.77$
				$p_2 = 0.23$, $\omega_2 = 3.39$			**$p_1 = 0.23$, $\omega = 3.39$**
genome	2574	$\omega = 0.41$	$p_0 = 0.74$, $\omega_0 = 0$	$p_0 = 0.84$, $\omega_0 = 0$	$p_0 = 0.92$, $\omega_0 = 0.10$	$B(0.03, 0.07)$	$B(0.04, 0.09)$
			$p_1 = 0.26$, $\omega_1 = 1$	$p_1 = 0.09$, $\omega_1 = 1$	**$p_1 = 0.08$, $\omega_1 = 2.82$**		$p_0 = 0.99$
				$p_2 = 0.02$, $\omega_2 = 8.88$	**$p_2 = 0.001$, $\omega_2 = 23.08$**		**$p_1 = 0.01$, $\omega = 9.45$**

Table 5.2 Likelihood ratio statistics (2 $\Delta\lambda$) for hypothesis testing

Gene	M0 vs. M3 ($\chi^2_{0.01,\,4} = 13.28$)	M1 vs. M2 ($\chi^2_{0.01,\,2} = 9.21$)	M7 vs. M8 ($\chi^2_{0.01,\,2} = 9.21$)
gag	24.90	10.54	13.08
pol	54.16	13.14	13.90
vif	25.18	9.42	9.56
env	200.02	97.98	98.26
nef	17.42	13.34	18.14
genome	331.94	105.92	105.58

Table 5.3 Sites identified as evolving by positive selection under M2

Gene	Single gene analysis	Genomic analysis
gag	24N, 146V	24N, 146V
pol	55A, **99V, 100A**, 171R, 251Y, 421N, 530T, 722I, 831S	**99V,** 100A, 421N, 722I
vif	27I, **29T**, 34T, **46S**,	29T, **34T**, 46S
env	131S, 137A, 140K, 143S, 145K, **147I**, 152S,!56T, 159A, 160E, 164V, **165H**, 211S, **368K**, 439T, **440A**, 441N, 444N, 447T, 449S	131S, 137A, 140K, 143S, 145K, **147I**, 164V, 165H, 211S, 368K, 439T, **440A**, **441N**, 444N, 447T, 449S
nef	8E, **14A**, **36T**, 43T, 45S, **69A**, 111T, 201L	14A, **36T**, 43T, 45S, **69A**,

5.3.2 Possible Recombination Detected in SIVmac Genomes

My results indicated that at least some of the sequence diversity observed was due to recombination. An ISI of 0.45 suggested that a moderate level of recombination has occurred in the genome of SIVmac. This result was found to be statistically significant with a probability score of less than 0.003. Similar estimates were obtained for the analyses of third codon positions and all codons indicating that my analyses were robust. I used PLATO to infer possible hybrid regions. As mentioned in Chapter 4, a Bonferroni-corrected significance (for multiple window size) was used to assess the results. It appeared to be that for an α of 0.05, any Z-value exceeded the threshold of 4.21was considered to be significant. Only one possible recombinant region was inferred with statistical significance.
At a Z-value of 6.98, region 5209-5239 was inferred to be a hybrid region. These nucleotides were mapped to *env*, where eleven candidate positive selection sites were located. These amino acids (1737-1747) were excluded from the results. To test the robustness of my findings, I excluded *env* from the dataset and repeated the analysis. No recombinant region was inferred once the gene was removed from the data.

5.3.3 Higher d_N/d_S Estimate for SIVmac Lineages in Comparison to SIVsm

Different ω estimates were obtained for SIVsm and SIVmac clades under model R2a and R3, indicating a non-uniform selective pattern across the phylogeny (Table 5.4). It appeared to be that SIVmac lineages have a much higher ω ratio (0.38), even when averaged across the

genome. LRT showed that Model R2a fitted the data much better than M0, implicating a statistically significant rate differences between the two clades (Table 5.5). Similarly large rate variation was observed between the internal branch and the rest of the tree (Table 5.5). Parameter estimates under Model R2b suggested a lower rate for the internal branch with the clades evolving at a faster rate (Table 5.4). Both two-ratio models (R2a and R2b) were rejected with significant statistical support when compared with the three-ratio model (R3). Model R3 estimated three different ω ratios for SIVmac (0.39), SIVsm (0.26) and internal lineage (0.07), which fitted the data much better, indicating that these rate differences must be accounted for. The ω estimate for SIVsm under R3 (0.26) was higher than under R2a (0.16) and this increase was supported by LRT (Table 5.5). As previously discussed, this low ω estimate could be produced by averaging the low internal branch rate (0.07) with the SIVsm clade (see Chapter 4). If selective pressure varied dramatically across the genome, averaging over all the genes would limit the power to detect lineage specific rate heterogeneity. Hence, I analysed four major protein coding genes *gag*, *pol*, *env* and *nef* to determine the extent of rate variation between genes and branches. As model R2a tends to underestimate ω for SIVsm lineages, it was not used in the single gene analysis. In all four genes, M0 was rejected when compared to Model R2b, indicating a low ω estimate for the internal branch (Table 5.4 and 5.5). For all the genes, the "three-ratio model" (R3) fitted the data significantly better than the "two-ratio model" (R2b). Parameter estimates indicated that all four genes were evolving at a higher rate in SIVmac, with notably the largest rate difference observed in *nef*. LRT detected divergent changes of selective constraint in *nef*.

Table 5.4 Parameter estimates under models allowing lineage specific evolution

Gene	One-ratio (M0)	Two-ratio (R2a)	Two-ratio (R2b)	Three-ratio (R3)
gag	$\omega = 0.13$	-	$\omega_{internal} = 0.05$	$\omega_{internal} = 0.05$
			$\omega_{clades} = 0.19$	$\boldsymbol{\omega_{SIVmac} = 0.32}$
				$\omega_{SIVsm} = 0.14$
pol	$\omega = 0.06$	-	$\omega_{internal} = 0.03$	$\omega_{internal} = 0.03$
			$\omega_{clades} = 0.11$	$\boldsymbol{\omega_{SIVmac} = 0.14}$
				$\omega_{SIVsm} = 0.09$
env	$\omega = 0.36$	-	$\omega_{internal} = 0.10$	$\omega_{internal} = 0.10$
			$\omega_{clades} = 0.54$	$\boldsymbol{\omega_{SIVmac} = 0.63}$
				$\omega_{SIVsm} = 0.48$
nef	$\omega = 0.43$	-	$\omega_{internal} = 0.15$	$\omega_{internal} = 0.15$
			$\omega_{clades} = 0.52$	$\boldsymbol{\omega_{SIVmac} = 1.09}$
				$\omega_{SIVsm} = 0.34$
genome	$\omega = 0.21$	$\omega_{SIVmac} = 0.38$	$\omega_{internal} = 0.07$	$\omega_{internal} = 0.07$
		$\omega_{SIVsm} = 0.16$	$\omega_{clades} = 0.30$	$\boldsymbol{\omega_{SIVmac} = 0.39}$
				$\omega_{SIVsm} = 0.26$

Table 5.5 Likelihood statistics (2 Δλ) for models of variable ω across the phylogeny

Gene	M0 vs. R2a ($\chi^2_{0.01, 1} = 6.63$)	M0 vs. R2b ($\chi^2_{0.01, 1} = 6.63$)	R3 vs. R2a ($\chi^2_{0.01, 1} = 6.63$)	R3 vs. R2b ($\chi^2_{0.01, 1} = 6.63$)
gag	-	22.00	-	8.80
pol	-	54.20	-	8.05
env	-	93.92	-	7.41
nef	-	10.62	-	14.02
genome	97.96	224.50	144.50	17.96

It appeared to be that a large proportion of the sites in the *nef* gene of SIVmac were evolving by adaptive evolution, with an averaged ω close to 1 (Table 5.4). This finding agreed with the site-specific analysis, where approximately 23% of the sites in *nef* were identified as positive selection sites (see Table 5.1). Interestingly, the substitution rate ratio remained less than one for SIVsm lineages ($\omega = 0.34$) indicating a change of selective constraint post zoonosis.

5.3.4 Positive Selection Detected in SIVmac and Across the Entire Phylogeny

Analyses of the datasets have shown that positive selections were detected on SIVmac and SIVsm lineages (Table 5.6). Models allowing among site rate variation (M3) have indicated that some sites were evolving by adaptive evolution across the entire phylogeny. Parameter estimates under M3 indicated that approximately 2% of the sites in SIVmac and SIVsm genomes were evolving by positive selection (Table 5.6).

Table 5.6 Parameter estimates under branch-site models

	Parameter estimates under different models					
Gene	M1 (neutral)	M3 (selection) $K = 2$	M3 (selection) $K = 3$	Model A (only allow selection in foreground lineages)	Model B (allows selection across the entire phylogeny)	Model D (allows two different rates for two different parts of the tree)
gag	-	$p_0 = 0.86$, $\omega_0 = 0.05$	$p_0 = 0.84$, $\omega_0 = 0.03$	-	$p_0 = 0.67$, $\omega_0 = 0.01$	$p_0 = 0.80$, $\omega_0 = 0.03$
		$p_2 = 0.14$, $\omega_2 = 1.36$	$p_1 = 0.15$, $\omega_1 = 0.71$		$p_1 = 0.11$, $\omega_1 = 0.92$	$p_1 = 0.18$, $\omega_1 = 0.53$
			$p_2 = 0.01$, $\omega_2 = 3.81$		$p_{(2+3)} = 0.22$, $\omega_2 = 0.81$	**$p_2 = 0.02$, $\omega_A = 1.35$, $\omega_B = 7.30$**
pol	-	$p_0 = 0.91$, $\omega_0 = 0.03$	$p_0 = 0.94$, $\omega_0 = 0$	-	$p_0 = 0.91$, $\omega_0 = 0.03$	$p_0 = 0.91$, $\omega_0 = 0.03$
		$p_2 = 0.09$, $\omega_2 = 1.03$	$p_1 = 0.05$, $\omega_1 = 1$		$p_1 = 0.06$, $\omega_1 = 0.68$	$p_1 = 0.05$, $\omega_1 = 0.96$
			$p_2 = 0.003$, $\omega_2 = 20.52$		**$p_{(2+3)} = 0.03$, $\omega_2 = 2.85$**	**$p_2 = 0.04$, $\omega_A = 0.21$, $\omega_B = 1.84$**
env	-	$p_0 = 0.85$, $\omega_0 = 0.08$	$p_0 = 0.72$, $\omega_0 = 0.01$	-	$p_0 = 0.81$, $\omega_0 = 0.06$	$p_0 = 0.72$, $\omega_0 = 0.00$
		$p_2 = 0.15$, $\omega_2 = 2.22$	$p_1 = 0.23$, $\omega_1 = 0.84$		**$p_1 = 0.16$, $\omega_1 = 1.81$**	**$p_1 = 0.05$, $\omega_1 = 5.88$**
			$p_2 = 0.05$, $\omega_2 = 5.88$		**$p_{(2+3)} = 0.03$, $\omega_2 = 6.74$**	**$p_2 = 0.23$, $\omega_A = 0.58$, $\omega_B = 1.19$**
nef	-	$p_0 = 0.85$, $\omega_0 = 0.14$	$p_0 = 0.56$, $\omega_0 = 0.00$	-	$p_0 = 0.74$, $\omega_0 = 0.10$	$p_0 = 0.56$, $\omega_0 = 0.00$
		$p_2 = 0.15$, $\omega_2 = 2.01$	$p_1 = 0.36$, $\omega_1 = 0.54$		**$p_1 = 0.12$, $\omega_1 = 2.29$**	**$p_1 = 0.05$, $\omega_1 = 4.06$**
			$p_2 = 0.08$, $\omega_2 = 2.86$		**$p_{(2+3)} = 0.14$, $\omega_2 = 2.11$**	**$p_2 = 0.39$, $\omega_A = 0.38$, $\omega_B = 1.94$**
genome	$p_0 = 0.79$, $\omega_0 = 0$	$p_0 = 0.84$, $\omega_0 = 0$	$p_0 = 0.81$, $\omega_0 = 0.03$	$p_0 = 0.72$, $\omega_0 = 0$	$p_0 = 0.84$, $\omega_0 = 0.05$	$p_0 = 0.74$, $\omega_0 = 0.02$
	$p_1 = 0.21$, $\omega_1 = 1$	**$p_2 = 0.02$, $\omega_2 = 8.88$**	$p_1 = 0.18$, $\omega_1 = 0.81$	$p_1 = 0.20$, $\omega_1 = 1$	**$p_1 = 0.09$, $\omega_1 = 1.93$**	**$p_1 = 0.03$, $\omega_1 = 4.58$**
			$p_2 = 0.02$, $\omega_2 = 4.59$	**$p_{(2+3)} = 0.08$, $\omega_2 = 3.09$**	**$p_{(2+3)} = 0.07$, $\omega_2 = 2.39$**	**$p_2 = 0.23$, $\omega_A = 0.43$, $\omega_B = 1.26$**

Table 5.7 Likelihood statistics (2 Δλ) for branch-site models

Gene	Model A vs. M1 ($\chi^2_{0.01,\,2} = 9.21$)	Model B vs. M3 ($K = 2$) ($\chi^2_{0.01,\,2} = 9.21$)	Model D vs. M3 ($K = 3$) ($\chi^2_{0.01,\,1} = 6.63$)
gag	-	77.28	7.70
pol	-	173.56	117.34
env	-	110.40	15.60
nef	-	114.82	25.30
genome	144.28	32.14	101.50

Averaging over the entire evolutionary history lacks the power to detect changes of selective pressure in different parts of the phylogeny. Hence, likelihood ratio statistics showed that branch-site models fitted the data significant better than their null, i.e. the site-specific models (Table 5.7). These models were able to detect lineage specific rate variation at certain amino acids. Both model A and B indicated a proportion of sites (~ 8 %) across the genome were evolving by positive selection in the foreground lineages. Model B also detected a fraction of positively selected sites across the entire phylogeny (~ 9% throughout the genome). Parameter estimates under model D suggested that roughly 3% of the sites were under positive selection throughout the evolutionary history (Table 5.6). Interestingly, divergent changes in selective constraint were detected in 23% of the sites across the genome. It appeared to be that these sites were evolving at a higher rate in SIVmac lineages (with ω_B = 1.26) than in SIVsm (ω_A = 0.43). At a threshold of 90%, the Bayesian method identified 33 positive selection sites across the genomes of SIVsm and SIVmac in the site-specific analysis of SIVsm/SIVmac dataset (See Table 5.8). The sites identified to be evolving by positive

selection in the foreground lineages (under model A and B) was a subset of the 33 sites identified in the site-specific analysis.

Table 5.8 Number of positive selection sites identified by M3, Model A, B and D.

Gene	M3	Model A	Model B		Model D	
	across the tree	foreground only	$p \geq 90\%$		$p \geq 90\%$	
	$p \geq 90\%$[a]	$p \geq 90\%$	foreground	across the tree	ω_A and ω_B[b]	across the tree
gag	2	-	0	0	5	0
pol	9	-	6	0	11	0
env	20	-	8	66	100	29
nef	8	-	4	13	53	8
Genome	33	24	22	57	151	34

[a] Positive selection sites identified under these models with posterior probabilities ≥ 90%.

[b] Model D detects divergent changes in selective constraint at certain amino acids, which might not be under adaptive evolution. Hence rate classes ω_A, and ω_B might contain sites that was not positively selected.

Model A did not appear to provide a significant improvement over site specific analyses, as it seemed to identify a subset of sites already detected by sites specific analyses. Hence, model A was not used in the subsequent single gene analyses. Parameter estimates under discrete models (M3; $K = 2$, and $K = 3$) suggested that each of the four major protein coding genes (*gag*, *pol*, *env*, and *nef*) possessed a proportion of sites evolving by positive selection across the entire phylogeny (Table 5.7). Estimates under branch-site models indicated that SIVmac genes and their progenitors were subjected to lineage specific changes in selective constraint.

The number of positive selection sites identified in each gene under branch-site models is presented in Table 5.8. Env and Nef appeared to be evolving under similar patterns of selective constraint, as both genes possessed a portion of sites evolving by positive selection in the foreground lineages, as well as across the entire phylogeny. The *gag* gene was more conserved in comparison, as model B was unable to detect adaptive evolution in any part of the tree (Table 5.6). However, divergent changes in selective constraint were detected in all the analysed genes. Comparisons of model D with its null (M3, $K = 3$) indicated that these selective changes were statistically significant (Table 5.7). In *pol*, *env*, and *nef*, a relatively large fraction of sites had a remarkable increase in ω in the foreground lineage. Most importantly, approximately 2% of the sites in *gag* were thought to be evolving by relaxed functional constraint in SIVsm, but have became positively selected in SIVmac lineages (see estimates under model D).

5.4 Discussion

5.4.1 Positive Selection Detected: Power and Accuracy

Successful detection of positive selection is dependent on the performance of site specific models, which in turn relies on the accuracy and power of likelihood ratio statistics. Simulation studies indicated that LRT is made conservative by comparing the rate distribution to a χ^2 distribution (Anisimova et al. 2001). Type I error rate is greatly reduced under the specified significance level of the test. Sequence length, divergence, intensity of

positive selection, and number of taxa sampled are factors influencing the power of LRT. Hence they must be taken into consideration when concluding positive selection. The sample size (i.e. number of sequences) seemed to be the most important factor, as the power to detect positive selection is greatly reduced for small datasets (Anisimova et al. 2001). In light of these studies, I conclude that strong positive selection must be operating throughout the genome of SIVmac, probably as a response to elevated selective pressure. The signal for positive selection was presumably strong, as the dataset was relatively small.

Previous statistical study on SIVmac genes suggested that adaptive evolution is only operating in highly immunogenic genes, such as *env* and *nef* (Shpaer and Mullins 1993). Purifying selection was thought to predominate in the more conserved genes, such as *pol*. Sequence divergence and rate variation observed in *gag* were attributed to relaxed functional constraint. My analyses indicated a non-uniform selective constraint was operating throughout the genome of SIVmac. Although purifying selection was acting on most of the sites in the genome, the remaining fraction was evolving by adaptive evolution, rather than relaxed functional constraint. Interestingly, this pattern of evolution differed from that of the human AIDS viruses (see chapter 2 and 4). Most protein coding genes in HIV-1 and 2 possessed a relatively large fraction of sites (~ 20%) evolving by relaxed functional constraint. In contrast, neutral evolution appeared to have little influence in the genetic diversity of SIVmac genes, as ω estimates was either close to 0 or greatly exceeded 1. As mentioned in chapter 2, the assumptions of M2 are highly unrealistic. Estimation under M2 showed that less than 10% of the sites in the genome were evolving "neutrally". Hence, even if selective constraint were relaxed at certain sites, it had occurred at fewer amino acids in SIVmac genes than in HIV genes. This could reflect a relatively recent introduction to the new hosts, as "neutral sites" tends to accumulate after a long period of relaxed functional

constraint. In fact this is a distinct possibility these macaques were experimentally infected and developed AIDS within a year; it is possible that most of the observed variations were either adaptive changes or escape mutations (Kuiken et al. 2001).

5.4.2 Adaptive Evolution Driven by CTL Recognition

There is a growing acceptance that CTL mediated viral clearance plays a critical role in controlling disease progression in pathogenic SIV infections (Kuroda et al 1999; Schmitz et al.1999; Allen et al. 2001; Vogel et al. 2002). As discussed in Chapter 4, a broad sustained CTL response is often observed in the asymptomatic infections of SIVsm, whereas the virulent infections of macaque is characterised by the decline of CTL activities. Hence, the evolution of CTL epitopes and escape mutations are considered to be means of evading immune detection. In the previous chapter, I presented evidence supporting the notion that CTL epitopes remained immunogenic in the natural host. However, these epitopes are unlikely to be recognised and targeted in a foreign host, due to differences in MHC-1 alleles. In fact, SIVmac specific epitopes differed markedly from SIVsm specific epitopes (O'Connor et al. 2001). SIVsm specific epitopes were thought to experience relaxed selective pressure post the cross-species transmission. In contrast, the new epitopes are now under intense immune selection. Many studies have reported an elevated fixation rate for replacement substitutions in some of these CTL epitopes (Courgnaud et al. 1998; Valli et al. 1998; Vogel et al. 2002). My analyses of the SIVmac genes assuming site-specific variation in selective constraint supported this notion. I have identified 46 candidate positive selection sites across the genome of SIVmac; eleven of these were located in the possible recombinant region and were excluded from epitope mapping. More than half of the remaining sites (~54%) were

located within or close to identified CTL epitopes. More importantly, positive selection sites identified in *gag*, and *nef* were also reported as CTL escape mutations in longitudinal experiments (Evans et al 1999; Chen et al. 2000;). It appeared to be that over half of the observed adaptive changes in SIVmac genome were driven by CTL recognitions. Recent study has suggested an active role for CD4+ cells in viral control (Horton et al. 2002). It is possible that the last 46% of the positive selection sites might be located in T- helper or antibody epitopes. However, one should exercise caution before reaching such conclusions, as relatively little is known regarding the location of these epitopes and their mode of evolution.

5.4.3 Post Zoonosis: Changes in Replacement Substitution Rates

Relatively little is known regarding the change of selective pressure post cross-species transmission. Although there is a unanimous agreement that the replacement substitution rates should increase post zoonosis (Hahn et al. 2000; Sharp et al. 2001), no comprehensive statistical study has yet been conducted to clarify this issue. This is partially due to the lack of data, as well as the late development of an accurate but powerful method. The development of codon models that allowed different ω ratios in different parts of the evolutionary history provide a mean for detect selective changes over time (Bielawski and Yang 2003). Although these lineage specific models estimate ω ratios as an average across all the sites, they are nonetheless a powerful method to detect rate variation in different time period. Using this approach, I have detected a significant rate difference between SIVmac and SIVsm lineages. A significant increase in the rate of amino acid replacements was observed since the arrival of SIVsm in macaques. It appeared to be that replacement

substitutions were fixed twice as fast in SIVmac as suppose to its progenitor. This rate increase could be due to adaptive evolution fixing advantageous changes, such as CTL escape mutations. Note that the largest rate increase is in the *nef* gene, where a strong signal for positive selection was detected. My findings suggested that amino acid substitutions in the SIVmac *nef* gene occur at a rate three times higher than in SIVsm. Interestingly, certain mutations in the SIVmac *nef* gene are associated with an increase of plasma viral load and a rapid development to disease (Kestler et al. 1991; Du et al. 1996; Kirchhoff et al. 1999). The expression of SIV and HIV Nef protein is also linked with the downregulation of CD4 and MHC-1 molecules. As previously discussed, a high level of *nef* specific CTL responses is observed at the acute phase of SIV infection of macaques. Hence, given the importance of *nef* in influencing pathogenicity and immune evasion, it is likely that it has undergone the most intense burst of evolution, when compared with the rest of the genome. To summarise, this study has provided statistical support for an accelerated rate of evolution post zoonosis.

5.4.4 Post Zoonosis: Divergent Changes in Selective Pressure at Specific Amino Acids

As discussed above, changes in selective pressure at certain amino acids are expected post a zoonotic transmission. Past study has shown that CTL epitope switching tends to occur after the cross-species jump (Kaur et al. 2001). Hence, sites that were under immune selection could be subjected to drift, as new epitopes are targeted. Likewise, amino acid replacement rates could increase at sites that were not previously selected. Identification of these sites would greatly enhance our understanding of cross species transmissions. Using the branch-site model developed by Yang and Nielsen (2002), I have detected adaptive evolution

operating at sites that were not under selection in the progenitor lineages. Model B was able to detect changes in selective pressure in most of the genes analysed, with the exception of *gag*. Approximately 38% of these positively selected sites across the genome were mapped to epitopes targeted by macaque CTLs. As expected none of these sites were mapped to regions recognised by mangabey CD8 + cells. Interestingly branch-site model B also identified a fraction of sites that were under continual selective pressure throughout the entire evolutionary history. These sites were located in highly immunogenic and fast evolving genes *env* and *nef*. At a threshold probability of 90%, Bayesian method has identified 66 and 13 sites evolving by continual selective pressure in *env* and *nef* respectively. In the *env* gene, 29 out of the 66 sites were mapped to regions recognised by CD8+ cells. Nine of these were located in epitopes targeted by macaque CTLs and the other ten were mapped in regions targeted in mangabeys (O' Connor et al. 2001). In the *nef* gene, five out of thirteen selected sites were mapped to epitopes identified by macaque CTLs. Hence, it appeared to be that at least half of the sites under continual selective pressure were associated with regions targeted by the immune system. Presumably, adaptive evolution at these sites promoted CTL escape.

As previously discussed these branch-site models may only detect episodic adaptive evolution, i.e. molecular adaptation occurred at certain sites for a short period of time (Bielawski and Yang et al. 2003). However, these models (model A and B) do not identify sites that were not evolving by adaptive evolution, even though divergent selective change might have occurred. Model D allows sites to evolve at two independent rates at different lineages and thus are able to detect changes in selective pressure that might not be promoted by positive selection. Divergent changes in selective constraint were detected in all the protein genes analysed (Table 5.6). In *nef*, a comparatively large fraction of sites (39%) that was evolving by relaxed functional constraint in the progenitor lineage became positively

selected after its arrival in macaque. More than half of these sites were mapped to macaque CTL epitopes, suggesting epitope switching might be responsible for this burst of adaptation. Similarly my findings indicated a two-fold increase in ω ratio at roughly 23% of the sites in *env* post zoonosis. Bayesian approaches have identified 100 sites scattered in *env*, 28 of these were located within a macaque CTL epitope and an additional 10 were mapped to mangabey epitopes. As the estimate for ω_B was close to one, it is likely that some of the sites identified in this rate class were evolving by near neutral evolution. It is possible that the 10 amino acids in mangabey epitope were "neutral sites", as new epitopes were targeted. In the *pol* gene, adaptive evolution was thought to be operating at eleven sites that were previously evolving by purifying selection ($\omega_A = 0.21$). Once again, over half of the sites were located in immunogenic regions of SIVmac indicating possible escape mutations. Most interestingly, model D detected dramatic changes in selective constraint at five sites in *gag*, which were not detected by model B. These sites (24N, 146V, 183T, 345V, and 382V) were estimated to have a ω close to one in the progenitor lineage ($\omega_A = 1.35$) indicating that synonymous changes were fixed at approximately the same rate as nonsynonymous changes. A significant increase in the rate of replacement substitutions was observed at these sites post zoonosis. As four of these sites (146V, 183T, 345V, and 382V) were mapped to the "new" CTL epitopes, this rate increase is likely to be promoted by immune recognition. However, model B identified these amino acids to be under relaxed functional constraint throughout the entire evolutionary history. The use of $\omega_0 \rightarrow \omega_2$ and $\omega_1 \rightarrow \omega_2$ rate classes, as well as the requirement of $\omega_2 > 1$ made model B conservative in detecting positive selection in the foreground lineages. It is possible that the conservative nature of this model also affected its power to detect changes in selective pressure.

5.4.5 Zoonosis Model Based on SIVmac Infection of Macaques

Many vaccine designs and studies of HIV pathogenesis are based on the SIVmac infections of macaques (Kuroda et al. 1999; Chen et al. 2000; Joag 2000; Horton et al. 2002; Vogel et al. 2002). As discussed in 5.1, many similarities and parallels between human and macaques, SIVmac and HIV, made the SIVmac/macaque an attractive model to study HIV-1 infection. It is possible that SIVmac/macaque could be used to model the zoonotic transmissions of SIVcpz to human, due to the lack of SIVcpz sequences. However, caution should be exercised in doing so, as my studies have shown difference between the evolution of HIV-2 and SIVmac post cross-species transmission. It appeared to be that SIVmac were evolving at a faster rate than HIV-2 post infection (as indicated by lineage specific studies). This apparent rate difference could contribute to the difference observed in pathogenicity. My results also indicated a fraction of sites that were evolving by relaxed functional constraint in SIVsm became positively selected in the foreign host. Half of these sites were located in CTL epitopes, indicating adaptive evolution driven by immune recognition. However, no such sites were identified in HIV-2; i.e. sites that were evolving by near neutral evolution did not become positively selected post cross-species transmission. Hence, it is possible that host dependent factors were involved in these observed differences. Future cross sectional and longitudinal studies of SIVcpz/HIV-1 is needed to elucidate the significance of changes in selective pressure post cross-species jump. However, the identification of sites undergone a burst of evolution post zoonosis was the first step in understanding adaptation in new hosts and changes in virulence.

CONCLUSION

The evolution of human AIDS viruses is a complex process that presented many interesting problems for different fields of science, ranging from experimental studies of virology to mathematical modelling. To elucidate the general patterns of evolution, one needs to unite experimental observations with mathematical predictions. In chapter 1, I covered the general background virology and briefly reviewed the origin and diversity of human AIDS viruses. In the recent years rapid advances in computer technology have greatly increased the computational powers that allowed us to test various theoretical methods. Theoretical modelling has proven to be a powerful and accurate tool in deciphering the complex issue of evolution. As demonstrated in my study, applications of maximum likelihood codon substitution models have helped to uncover many facets of HIV evolution.

In chapter 2, I explored the idea of approximating rates of substitution into statistical distributions and hence allowing among site rate variation. This approach has revealed that all the sites across the genome of HIV-1 were evolving by different intensities of selective constraint. I was able to identify regions of functional importance, judging by the amino acid conservation. Most excitingly, I have detected that each gene possessed a fraction of sites evolving by positive Darwinian selection. Most of these sites were mapped to known immunogenic epitopes indicating adaptive evolution driven by immune recognition. In the later half of this chapter, I explored the acceptability of these positive selection sites. Acceptability was defined as the ability to accommodate residues of diverse physiochemical property. My results indicated that positive selection appeared to operate within structural constraints. Physiochemically conserved amino acid changes were observed at buried sites, whereas radical changes predominated at exposed residues. These were the first two general

patterns of HIV-1 evolution I have uncovered in my study and they warranted further investigation.

The third chapter was an extension on the findings of chapter 2. In the first section, I examined the possibility of subtype-specific variations in selective pressure and their implications. I analysed 55 genomic sequences of subtype B and 40 of subtype C. As many vaccine studies were based on subtype B infections, which are most prevalent in USA, it is imperative that these vaccines are effective against subtype C infections from Africa. Goschen et al. (2002) has reported subtype-specific variation of selective pressure in the *env* gene of subtype B and C. I added to this observation by noting that most HIV-1 genes were evolving at a subtype specific manner. Subtype-specific variations in selective constraint existed not only in *env*, but it could be observed across the entire genome. As adaptive evolution is driven chiefly by immune detection, this change of selective constraint is indicative of variations in immune targeting. Such differences in selective pressure contributed to the extensive genetic diversity observed and could be viewed as co-evolution of the subtypes. This is the third general pattern of evolution observed in HIV-1. The second part of chapter 3 illustrated the amino acid substitution pattern with respect to structural constraint and physiochemical properties. In general, the rates of conservative substitutions greatly exceeded that of radical substitutions. Intense purifying selection was thought to reduce the rate at which radical changes were fixed. However, a different pattern was observed at positive selection sites. It appeared to be that positive selection promoted both conservative and radical substitutions, whereas relaxed functional constraint only resulted in conserved changes. This is notion is consistent with the observations in chapter 2, as I pooled both buried and exposed positive selection sites into one large dataset as a test for robustness. Interestingly, mutations of certain physiochemical properties were more constrained than the

others. For example, changes involving alterations of both polarity and volume of a residue were heavily selected against, whereas changes in polarity alone were better tolerated. The structural location of a residue also seemed to affect the rate of amino acid substitution. As discussed in chapter 2, amino acid replacements at buried residues tend to be physiochemically conservative. In this chapter, I discovered that at buried residues, conservative mutations (with respect to charge) were under similar intensity of purifying selection as radical substitutions. Hence, the fourth general trend of HIV evolution was that patterns of amino acid substitution were influenced by positive selection, physiochemical properties and structural requirements.

One of the hypotheses regarding the evolution of human AIDS viruses is that a burst of evolution must have occurred immediately after their arrival in human. However, the evolution of HIV-1 immediately post zoonosis remained difficult to elucidate, as few sequences of the progenitor virus (SIVcpz) were available. Hence, I shifted my focus in chapter 4 and examined the evolutionary changes of HIV-2 post the cross-species transmission. As HIV-2 infections were associated with slower disease progression and a lower virulence, one could expect a slightly different pattern of evolution at the molecular level. My findings have shown that although adaptive evolution was detected across the genomes of HIV-2, fewer positive selection sites were identified as compared with HIV-1. Interestingly, positive selection was also detected in the progenitor virus that remained avirulent in sooty mangabey. However, only the *env* gene appeared to be evolving by adaptive evolution in SIVsm, indicating a sustained immunogenicity even in the natural host. My results also indicated that HIV-2 lineages do not appeared to be evolving at a faster rate than SIVsm branches. It is possible that fewer adaptive changes in the progenitor virus were required for successful infection.

HIV-2 and SIVmac both arose from the same progenitor, yet their virulence differed dramatically. Macaques infected with SIVmac showed rapid progression to AIDS and in extreme cases died within days. SIVmac/macaque has become a popular primate model for the study of pathogenesis, due to certain similarities it shares with the HIV-1 infections. In the last chapter, I explored the possibility of using the popular primate model (SIVmac/macaque) to describe the zoonosis of human AIDS viruses. My findings suggest that post zoonosis; SIVmac appeared to be evolving much faster than HIV-2. As the macaques were likely to be experimentally infected, any changes I have observed are adaptive changes. My results have shown that a fraction of sites that were evolving by relaxed functional constraint in the progenitor lineage have undergone a burst of evolution, since its arrival in macaque. This sudden divergent change in selective pressure was found to be associated with epitope switching. Given that the evolution of SIVmac post zoonosis differed dramatically from that of HIV-2, it is perhaps unwise to use SIVmac/macaque as an evolutionary model for AIDS viruses.

I have set out to explore the evolution of human AIDS viruses and I have gained some insight into this complex process. However, I have raised more questions than I have tried to answer. How does one apply the trends of evolution to vaccine design? How can this information be used to eradicate AIDS? Exactly what changes pathogenesis? To what extent is the virulence dependent on host? What will become of the subtypes in the next decade? What would HIV-1 evolve into? The satisfactory answers to these questions may not be revealed for a long time. However, one hopes that they will be answered in the years to come.

LITERATURE CITATION

Alff-Steinberger, C. 1969. The genetic code and error transmission. *Proc.Natl.Acad.Sci.U.S.A* 64:584-591.

Allen, T. M., D. H. O'Connor, P. Jing, J. L. Dzuris, B. R. Mothe, T. U. Vogel, E. Dunphy, M. E. Liebl, C. Emerson, N. Wilson, K. J. Kunstman, X. Wang, D. B. Allison, A. L. Hughes, R. C. Desrosiers, J. D. Altman, S. M. Wolinsky, A. Sette, and D. I. Watkins. 2000. Tat-specific cytotoxic T lymphocytes select for SIV escape variants during resolution of primary viraemia. *Nature* 407:386-390.

Allen, T. M., B. R. Mothe, J. Sidney, P. Jing, J. L. Dzuris, M. E. Liebl, T. U. Vogel, D. H. O'Connor, X. Wang, M. C. Wussow, J. A. Thomson, J. D. Altman, D. I. Watkins, and A. Sette. 2001. CD8(+) lymphocytes from simian immunodeficiency virus-infected rhesus macaques recognize 14 different epitopes bound by the major histocompatibility complex class I molecule mamu-A*01: implications for vaccine design and testing. *J.Virol.* 75:738-749.

Altfeld, M., E. S. Rosenberg, R. Shankarappa, J. S. Mukherjee, F. M. Hecht, R. L. Eldridge, M. M. Addo, S. H. Poon, M. N. Phillips, G. K. Robbins, P. E. Sax, S. Boswell, J. O. Kahn, C. Brander, P. J. Goulder, J. A. Levy, J. I. Mullins, and B. D. Walker. 2001. Cellular immune responses and viral diversity in individuals treated during acute and early HIV-1 infection. *J.Exp.Med.* 193:169-180.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol.Biol.Evol.* 18:1585-1592.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol.Biol.Evol.* 19:950-958.

Anisimova, M., R. Nielsen, and Z. Yang. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164:1229-1236.

Barre-Sinoussi, F., J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* 220:868-871.

Beaumont, T., A. van Nuenen, S. Broersen, W. A. Blattner, V. V. Lukashov, and H. Schuitemaker. 2001. Reversal of human immunodeficiency virus type 1 IIIB to a neutralization-resistant phenotype in an accidentally infected laboratory worker with a progressive clinical course. *J.Virol.* 75:2246-2252.

Beer BE, Bailes E, Sharp PM, Hirsch VM (1999). Diversity and evolution of primate lentiviruses. pp. 460-474 in *Human Retroviruses and AIDS 1999*. Edited by: Kuiken CL, Foley B, Hahn B, Korber B, McCutchan F, Marx PA, Mellors JW, Mullins JI, Sodroski J, and Wolinksy S. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

Bielawski, J. P. and Z. Yang. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. *J.Struct.Funct.Genomics* 3:201-212.

Bondada S, Chelvarajan RL. 2001. B lymphocytes. *Encyclopedia of Science 2001.* Nature Publishing Group.

Calarota, S., M. Jansson, M. Levi, K. Broliden, O. Libonatti, H. Wigzell, and B. Wahren. 1996. Immunodominant glycoprotein 41 epitope identified by seroreactivity in HIV type 1-infected individuals. *AIDS Res.Hum.Retroviruses* 12:705-713.

Chang, L. J., C. H. Chen, V. Urlacher, and T. Z. Lee. 2000. Differential apoptosis effects of primate lentiviral Vpr and Vpx in mammalian cells. *J.Biomed.Sci.* 7:322-333.

Charleston, M. A. and D. L. Robertson. 2002. Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Syst.Biol.* 51:528-535.

Chen, Z. W., A. Craiu, L. Shen, M. J. Kuroda, U. C. Iroku, D. I. Watkins, G. Voss, and N. L. Letvin. 2000. Simian immunodeficiency virus evades a dominant epitope-specific cytotoxic T lymphocyte response through a mutation resulting in the accelerated dissociation of viral peptide and MHC class I. *J.Immunol.* 164:6474-6479.

Corbet, S., M. C. Muller-Trutwin, P. Versmisse, S. Delarue, A. Ayouba, J. Lewis, S. Brunak, P. Martin, F. Brun-Vezinet, F. Simon, F. Barre-Sinoussi, and P. Mauclere. 2000. env sequences of simian immunodeficiency viruses from chimpanzees in Cameroon are strongly related to those of human immunodeficiency virus group N from the same geographic area. *J.Virol.* 74:529-534.

Courgnaud, V., W. Saurin, F. Villinger, and P. Sonigo. 1998. Different evolution of simian immunodeficiency virus in a natural host and a new host. *Virology* 247:41-50.

Courgnaud, V., X. Pourrut, F. Bibollet-Ruche, E. Mpoudi-Ngole, A. Bourgeois, E. Delaporte, and M. Peeters. 2001. Characterization of a novel simian immunodeficiency virus from guereza colobus monkeys (Colobus guereza) in Cameroon: a new lineage in the nonhuman primate lentivirus family. *J.Virol.* 75:857-866.

Crandall, K. A., C. R. Kelsey, H. Imamichi, H. C. Lane, and N. P. Salzman. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol.Biol.Evol.* 16:372-382.

da Silva, J. and A. L. Hughes. 1998. Conservation of cytotoxic T lymphocyte (CTL) epitopes as a host strategy to constrain parasite adaptation: evidence from the nef gene of human immunodeficiency virus 1 (HIV-1). *Mol.Biol.Evol.* 15:1259-1268.

Dong, X. N., Y. Xiao, M. P. Dierich, and Y. H. Chen. 2001. N- and C-domains of HIV-1 gp41: mutation, structure and functions. *Immunol.Lett.* 75:215-220.

Doolittle, R. F. 1989. Immunodeficiency viruses: the simian-human connection. *Nature* 339:338-339.

Du, Z., P. O. Ilyinskii, V. G. Sasseville, M. Newstein, A. A. Lackner, and R. C. Desrosiers. 1996. Requirements for lymphocyte activation by unusual strains of simian immunodeficiency virus. *J.Virol.* 70:4157-4161.

Dzuris, J. L., J. Sidney, E. Appella, R. W. Chesnut, D. I. Watkins, and A. Sette. 2000. Conserved MHC class I peptide binding motif between humans and rhesus macaques. *J.Immunol.* 164:283-291.

Esteves, A., J. Piedade, C. Santos, T. Venenno, W. F. Canas-Ferreira, and R. Parreira. 2001. Follow-up study of intrahost HIV type 2 variability reveals discontinuous evolution of C2V3 sequences. *AIDS Res.Hum.Retroviruses* 17:253-256.

Evans, D. T., D. H. O'Connor, P. Jing, J. L. Dzuris, J. Sidney, J. da Silva, T. M. Allen, H. Horton, J. E. Venham, R. A. Rudersdorf, T. Vogel, C. D. Pauza, R. E. Bontrop, R. DeMars, A. Sette, A. L. Hughes, and D. I. Watkins. 1999. Virus-specific cytotoxic T-lymphocyte responses select for amino-acid variation in simian immunodeficiency virus Env and Nef. *Nat.Med.* 5:1270-1276.

Eyre-Walker, A., P. D. Keightley, N. G. Smith, and D. Gaffney. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol.Biol.Evol.* 19:2142-2149.

Fares, M. A., A. Moya, C. Escarmis, E. Baranowski, E. Domingo, and E. Barrio. 2001. Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (FMDV) subjected to experimental passage regimens. *Mol.Biol.Evol.* 18:10-21.

Fischer, W. B. and M. S. Sansom. 2002. Viral ion channels: structure and function. *Biochim.Biophys.Acta* 1561:27-45.

Foley, B. T. 2000. An overview of the molecular phylogeny of lentiviruses. pp. 35-43 in *HIV Sequence Compendium 2000*. Edited by: Kuiken C, McCutchan F, Foley B, Mellors JW, Hahn B, Mullins J, Marx P, Wolinsky S. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

Fomsgaard A. 1999 HIV-1 DNA vaccines. *Immunol Lett* 65: 127-131

Fu, Y. X. 2001. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. *Mol.Biol.Evol.* 18:620-626.

Gao, F., E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn. 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* 397:436-441.

Gaschen, B., J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. H. Hahn, T. Bhattacharya, and B. Korber. 2002. Diversity considerations in HIV-1 vaccine selection. *Science* 296:2354-2360.

Gojobori, T., E. N. Moriyama, Y. Ina, K. Ikeo, T. Miura, H. Tsujimoto, M. Hayami, and S. Yokoyama. 1990. Evolutionary origin of human and simian immunodeficiency viruses. *Proc.Natl.Acad.Sci.U.S.A* 87:4108-4111.

Gojobori, T., E. N. Moriyama, and M. Kimura. 1990. Molecular clock of viral evolution, and the neutral theory. *Proc.Natl.Acad.Sci.U.S.A* 87:10015-10018.

Gotch FM. 2001 T Lymphocytes: Cytotoxic. *Encyclopedia of Science 2001*. Nature Publishing Group.

Goudsmit, J. and V. V. Lukashov. 1999. Dating the origin of HIV-1 subtypes. *Nature* 400:325-326.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185:862-864.

Grassly, N. C. and E. C. Holmes. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol.Biol.Evol.* 14:239-247.

Hahn, B. H., G. M. Shaw, K. M. De Cock, and P. M. Sharp. 2000. AIDS as a zoonosis: scientific and public health implications. *Science* 287:607-614.

Harcourt, G. C., S. Garrard, M. P. Davenport, A. Edwards, and R. E. Phillips. 1998. HIV-1 variation diminishes CD4 T lymphocyte recognition. *J.Exp.Med.* 188:1785-1793.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J.Mol.Evol.* 22:160-174.

Haydon, D., S. Lea, L. Fry, N. Knowles, A. R. Samuel, D. Stuart, and M. E. Woolhouse. 1998. Characterizing sequence variation in the VP1 capsid proteins of foot and mouth disease virus (serotype 0) with respect to virion structure. *J.Mol.Evol.* 46:465-475.

Haydon, D. T., A. D. Bastos, N. J. Knowles, and A. R. Samuel. 2001. Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates. *Genetics* 157:7-15.

Hill, C. P., D. Worthylake, D. P. Bancroft, A. M. Christensen, and W. I. Sundquist. 1996. Crystal structures of the trimeric human immunodeficiency virus type 1 matrix protein: implications for membrane association and assembly. *Proc.Natl.Acad.Sci.U.S.A* 93:3099-3104.

Hirsch, V. M., G. Dapolito, C. McGann, R. A. Olmsted, R. H. Purcell, and P. R. Johnson. 1989. Molecular cloning of SIV from sooty mangabey monkeys. *J.Med.Primatol.* 18:279-285.

Holmes, E. C. 2001. On the origin and evolution of the human immunodeficiency virus (HIV). *Biol.Rev.Camb.Philos.Soc.* 76:239-254.

Horton, H., T. Vogel, D. O'Connor, L. Picker, and D. I. Watkins. 2002. Analysis of the immune response and viral evolution during the acute phase of SIV infection. *Vaccine* 20:1927-1932.

Horuk, R. 2001. Chemokine receptors. *Cytokine Growth Factor Rev.* 12:313-335.

Hughes, A. L., T. Ota, and M. Nei. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Mol.Biol.Evol.* 7:515-524.

Hughes, A. L. 1992. Positive selection and interallelic recombination at the merozoite surface antigen-1 (MSA-1) locus of Plasmodium falciparum. *Mol.Biol.Evol.* 9:381-393.

Hughes, A. L., M. K. Hughes, C. Y. Howell, and M. Nei. 1994. Natural selection at the class II major histocompatibility complex loci of mammals. *Philos.Trans.R.Soc.Lond B Biol.Sci.* 346:359-366.

Hughes, M. K. and A. L. Hughes. 1993. Evolution of duplicate genes in a tetraploid animal, Xenopus laevis. *Mol.Biol.Evol.* 10:1360-1369.

Hunt, R. D., B. J. Blake, L. V. Chalifoux, P. K. Sehgal, N. W. King, and N. L. Letvin. 1983. Transmission of naturally occurring lymphoma in macaque monkeys. *Proc.Natl.Acad.Sci.U.S.A* 80:5085-5089.

Igarashi, T., T. Kuwata, J. Takehisa, K. Ibuki, R. Shibata, R. Mukai, T. Komatsu, A. Adachi, E. Ido, and M. Hayami. 1996. Genomic and biological alteration of a human immunodeficiency virus type 1 (HIV-1)-simian immunodeficiency virus strain mac chimera, with HIV-1 Env, recovered from a long-term carrier monkey. *J.Gen.Virol.* 77 (Pt 8):1649-1658.

Jetzt, A. E., H. Yu, G. J. Klarmann, Y. Ron, B. D. Preston, and J. P. Dougherty. 2000. High rate of recombination throughout the human immunodeficiency virus type 1 genome. *J.Virol.* 74:1234-1240.

Joag, S. V. 2000. Primate models of AIDS. *Microbes.Infect.* 2:223-229.

Kaur, A., J. Yang, D. Hempel, L. Gritz, G. P. Mazzara, H. McClure, and R. P. Johnson. 2000. Identification of multiple simian immunodeficiency virus (SIV)-specific CTL epitopes in sooty mangabeys with natural and experimentally acquired SIV infection. *J.Immunol.* 164:934-943.

Kaur, A., L. Alexander, S. I. Staprans, L. Denekamp, C. L. Hale, H. M. McClure, M. B. Feinberg, R. C. Desrosiers, and R. P. Johnson. 2001. Emergence of cytotoxic T

lymphocyte escape mutations in nonpathogenic simian immunodeficiency virus infection. *Eur.J.Immunol.* 31:3207-3217.

Kestler, H. W., III, D. J. Ringler, K. Mori, D. L. Panicali, P. K. Sehgal, M. D. Daniel, and R. C. Desrosiers. 1991. Importance of the nef gene for maintenance of high virus loads and for development of AIDS. *Cell* 65:651-662.

Kirchhoff, F., S. Carl, S. Sopper, U. Sauermann, K. Matz-Rensing, and C. Stahl-Hennig. 1999. Selection of the R17Y substitution in SIVmac239 nef coincided with a dramatic increase in plasma viremia and rapid progression to death. *Virology* 254:61-70.

Kuiken C. 2001. Reagents for HIV/SIV Vaccine Studies. pp. 182-190 in *HIV Sequence Compendium 2001*. Edited by: Kuiken C, Foley B, Hahn B, Marx P, McCutchan F, Mellors JW, Wolinsky S, Korber B. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, LA-UR 02-2877.

Kuroda, M. J., J. E. Schmitz, W. A. Charini, C. E. Nickerson, M. A. Lifton, C. I. Lord, M. A. Forman, and N. L. Letvin. 1999. Emergence of CTL coincides with clearance of virus during primary simian immunodeficiency virus infection in rhesus monkeys. *J.Immunol.* 162:5127-5133.

Kyte, J. and R. F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J.Mol.Biol.* 157:105-132.

Langedijk, J. P., G. Zwart, J. Goudsmit, and R. H. Meloen. 1995. Fine specificity of antibody recognition may predict amino acid substitution in the third variable region of gp120 during HIV type 1 infection. *AIDS Res.Hum.Retroviruses* 11:1153-1162.

Letvin, N. L., K. A. Eaton, W. R. Aldrich, P. K. Sehgal, B. J. Blake, S. F. Schlossman, N. W. King, and R. D. Hunt. 1983. Acquired immunodeficiency syndrome in a colony of macaque monkeys. *Proc.Natl.Acad.Sci.U.S.A* 80:2718-2722.

Lukashov, V. V. and J. Goudsmit. 1997. Evolution of the human immunodeficiency virus type 1 subtype-specific V3 domain is confined to a sequence space with a fixed distance to the subtype consensus. *J.Virol.* 71:6332-6338.

McClellan, D. A. and K. G. McCracken. 2001. Estimating the influence of selection on the variable amino acid sites of the cytochrome B protein functional domains. *Mol.Biol.Evol.* 18:917-925.

McLain, L., J. L. Brown, L. Cheung, S. A. Reading, C. Parry, T. D. Jones, S. M. Cleveland, and N. J. Dimmock. 2001. Different effects of a single amino acid substitution on three adjacent epitopes in the gp41 C-terminal tail of a neutralizing antibody escape mutant of human immunodeficiency virus type 1. *Arch.Virol.* 146:157-166.

McVean, G., P. Awadalla, and P. Fearnhead. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160:1231-1241.

Miyata, T., S. Miyazawa, and T. Yasunaga. 1979. Two types of amino acid substitutions in protein evolution. *J.Mol.Evol.* 12:219-236.

Muller-Trutwin, M. C., S. Corbet, M. D. Tavares, V. M. Herve, E. Nerrienet, M. C. Georges-Courbot, W. Saurin, P. Sonigo, and F. Barre-Sinoussi. 1996. The evolutionary rate of nonpathogenic simian immunodeficiency virus (SIVagm) is in agreement with a rapid and continuous replication in vivo. *Virology* 223:89-102.

Nielsen, R. and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929-936.

Norris, P. J., M. Sumaroka, C. Brander, H. F. Moffett, S. L. Boswell, T. Nguyen, Y. Sykulev, B. D. Walker, and E. S. Rosenberg. 2001. Multiple effector functions mediated by human immunodeficiency virus-specific CD4(+) T-cell clones. *J.Virol.* 75:9771-9779.

Novembre, F. J, 2001. Simian Retroviruses. *Encyclopedia of Science 2001*. Nature Publishing Group.

O'Connor, D., T. Friedrich, A. Hughes, T. M. Allen, and D. Watkins. 2001. Understanding cytotoxic T-lymphocyte escape during simian immunodeficiency virus infection. *Immunol.Rev.* 183:115-126.

Onanga, R., C. Kornfeld, I. Pandrea, J. Estaquier, S. Souquiere, P. Rouquet, V. P. Mavoungou, O. Bourry, S. M'Boup, F. Barre-Sinoussi, F. Simon, C. Apetrei, P. Roques, and M. C. Muller-Trutwin. 2002. High levels of viral replication contrast with only transient changes in CD4(+) and CD8(+) cell numbers during the early phase of experimental infection with simian immunodeficiency virus SIVmnd-1 in Mandrillus sphinx. *J.Virol.* 76:10256-10263.

Overbaugh, J. and C. R. Bangham. 2001. Selection forces and constraints on retroviral sequence variation. *Science* 292:1106-1109.

Pedersen, F. S, and M, Duch. 2001. Retroviruses in Human Gene Therapy. *Encyclopaedia of Science 2001*. Nature Publishing Group.

Peek, A. S., V. Souza, L. E. Eguiarte, and B. S. Gaut. 2001. The interaction of protein structure, selection, and recombination on the evolution of the type-1 fimbrial major subunit (fimA) from Escherichia coli. *J.Mol.Evol.* 52:193-204.

Phillips, R. E., G. C. Harcourt, and D. A. Price. 2001. CD4+ T cells: The great escape. *Nat.Med.* 7:777-778.

Plikat, U., K. Nieselt-Struwe, and A. Meyerhans. 1997. Genetic drift can dominate short-term human immunodeficiency virus type 1 nef quasispecies evolution in vivo. *J.Virol.* 71:4233-4240.

Rey-Cuille, M. A., J. L. Berthier, M. C. Bomsel-Demontoy, Y. Chaduc, L. Montagnier, A. G. Hovanessian, and L. A. Chakrabarti. 1998. Simian immunodeficiency virus replicates to high levels in sooty mangabeys without inducing disease. *J.Virol.* 72:3872-3886.

Robertson, D. L., P. M. Sharp, F. E. McCutchan, and B. H. Hahn. 1995. Recombination in HIV-1. *Nature* 374:124-126.

Rosenberg, E. S., M. Altfeld, S. H. Poon, M. N. Phillips, B. M. Wilkes, R. L. Eldridge, G. K. Robbins, R. T. D'Aquila, P. J. Goulder, and B. D. Walker. 2000. Immune control of HIV-1 after early treatment of acute infection. *Nature* 407:523-526.

Ross, H. A. and A. G. Rodrigo. 2002. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *J.Virol.* 76:11715-11720.

Schmitz, J. E., M. A. Lifton, K. A. Reimann, D. C. Montefiori, L. Shen, P. Racz, K. Tenner-Racz, M. W. Ollert, M. A. Forman, R. S. Gelman, C. W. Vogel, and N. L. Letvin. 1999. Effect of complement consumption by cobra venom factor on the the course of primary infection with simian immunodeficiency virus in rhesus monkeys. *AIDS Res.Hum.Retroviruses* 15:195-202.

Seibert, S. A., C. Y. Howell, M. K. Hughes, and A. L. Hughes. 1995. Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (HIV-1). *Mol.Biol.Evol.* 12:803-813.

Sharp, P. M. 1997. In search of molecular darwinism. *Nature* 385:111-112.

Sharp, P. M., E. Bailes, R. R. Chaudhuri, C. M. Rodenburg, M. O. Santiago, and B. H. Hahn. 2001. The origins of acquired immune deficiency syndrome viruses: where and when? *Philos.Trans.R.Soc.Lond B Biol.Sci.* 356:867-876.

Sharp, P. M. 2002. Origins of human virus diversity. *Cell* 108:305-312.

Shpaer, E. G. and J. I. Mullins. 1993. Rates of amino acid change in the envelope protein correlate with pathogenicity of primate lentiviruses. *J.Mol.Evol.* 37:57-65.

Siliciano RFS 2001 Acquired Immune Deficiency Syndrome (AIDS). Encyclopedia of Science 2001. Nature Publishing Group.

Souquiere, S., F. Bibollet-Ruche, D. L. Robertson, M. Makuwa, C. Apetrei, R. Onanga, C. Kornfeld, J. C. Plantier, F. Gao, K. Abernethy, L. J. White, W. Karesh, P. Telfer, E. J. Wickings, P. Mauclere, P. A. Marx, F. Barre-Sinoussi, B. H. Hahn, M. C. Muller-Trutwin, and F. Simon. 2001. Wild Mandrillus sphinx are carriers of two types of lentivirus. *J.Virol.* 75:7086-7096.

Swanson, W. J., Z. Yang, M. F. Wolfner, and C. F. Aquadro. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc.Natl.Acad.Sci.U.S.A* 98:2509-2514.

Takehisa, J., Y. Harada, N. Ndembi, I. Mboudjeka, Y. Taniguchi, C. Ngansop, S. Kuate, L. Zekeng, K. Ibuki, T. Shimada, B. Bikandou, Y. Yamaguchi-Kabata, T. Miura, M. Ikeda, H. Ichimura, L. Kaptue, and M. Hayami. 2001. Natural infection of wild-born mandrills (Mandrillus sphinx) with two different types of simian immunodeficiency virus. *AIDS Res.Hum.Retroviruses* 17:1143-1154.

Taylor, W. R. 1986. The classification of amino acid conservation. *J.Theor.Biol.* 119:205-218.

Valli, P. J. and J. Goudsmit. 1998. Structured-tree topology and adaptive evolution of the simian immunodeficiency virus SIVsm envelope during serial passage in rhesus macaques according to likelihood mapping and quartet puzzling. *J.Virol.* 72:3673-3683.

van Rensburg, E. J., S. Engelbrecht, J. Mwenda, J. D. Laten, B. A. Robson, T. Stander, and G. K. Chege. 1998. Simian immunodeficiency viruses (SIVs) from eastern and southern Africa: detection of a SIVagm variant from a chacma baboon. *J.Gen.Virol.* 79 (Pt 7):1809-1814.

Vogel, T. U., T. C. Friedrich, D. H. O'Connor, W. Rehrauer, E. J. Dodds, H. Hickman, W. Hildebrand, J. Sidney, A. Sette, A. Hughes, H. Horton, K. Vielhuber, R. Rudersdorf, I. P. De Souza, M. R. Reynolds, T. M. Allen, N. Wilson, and D. I. Watkins. 2002. Escape in one of two cytotoxic T-lymphocyte epitopes bound by a high-frequency major histocompatibility complex class I molecule, Mamu-A*02: a paradigm for virus evolution and persistence? *J.Virol.* 76:11623-11636.

Walker, B. D. and P. J. Goulder. 2000. AIDS. Escape from the immune system. *Nature* 407:313-314.

Walther D. 1997 WebMol - a Java based PDB viewer. *Trends Biochem Sci*, 22: 274-275

Worobey, M. 2001. A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. *Mol.Biol.Evol.* 18:1425-1434.

Yamaguchi-Kabata, Y. and T. Gojobori. 2000. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J.Virol.* 74:4335-4350.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol.Biol.Evol.* 15:568-573.

Yang, Z., R. Nielsen, and M. Hasegawa. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol.Biol.Evol.* 15:1600-1611.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431-449.

Yang, Z. 2001. Maximum likelihood analysis of adaptive evolution in HIV-1 gp120 env gene. *Pac.Symp.Biocomput.*226-237.

Yang, Z. and R. Nielsen. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol.Biol.Evol.* 19:908-917.

Yang, Z. and W. J. Swanson. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol.Biol.Evol.* 19:49-57.

Zanotto, P. M., E. G. Kallas, R. F. de Souza, and E. C. Holmes. 1999. Genealogical evidence for positive selection in the nef gene of HIV-1. *Genetics* 153:1077-1089.

Zhang, J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J.Mol.Evol.* 50:56-68.

Zhang, P. F., X. Chen, D. W. Fu, J. B. Margolick, and G. V. Quinnan, Jr. 1999. Primary virus envelope cross-reactivity of the broadening neutralizing antibody response during early chronic human immunodeficiency virus type 1 infection. *J.Virol.* 73:5225-5230.